Microbial life cycles link global modularity in regulation to mosaic evolution

van Gestel, Jordi ; Ackermann, Martin ; Wagner, Andreas

Abstract: Microbes are exposed to changing environments, to which they can respond by adopting various lifestyles such as swimming, colony formation or dormancy. These lifestyles are often studied in isolation, thereby giving a fragmented view of the life cycle as a whole. Here, we study lifestyles in the context of this whole. We first use machine learning to reconstruct the expression changes underlying life cycle progression in the bacterium Bacillus subtilis, based on hundreds of previously acquired expression profiles. This yields a timeline that reveals the modular organization of the life cycle. By analysing over 380 Bacillales genomes, we then show that life cycle modularity gives rise to mosaic evolution in which life stages such as motility and sporulation are conserved and lost as discrete units. We postulate that this mosaic conservation pattern results from habitat changes that make these life stages obsolete or detrimental. Indeed, when evolving eight distinct Bacillales strains and species under laboratory conditions that favour colony growth, we observe rapid and parallel losses of the sporulation life stage across species, induced by mutations that affect the same global regulator. We conclude that a life cycle perspective is pivotal to understanding the causes and consequences of modularity in both regulation and evolution.

DOI: https://doi.org/10.1038/s41559-019-0939-6

Microbial life cycles link global modularity in regulation to mosaic evolution

Supplementary information

van Gestel, J., Ackermann, M., Wagner, A.

Supplementary Text 1. Brief overview of reconstructions of global transcription network of *Bacillus subtilis*.

Studies on the global regulatory network of *B. subtilis* can generally be divided in two categories: (1) *descriptive* studies[1–4], which aim to characterize the main properties of the network, and (2) *exploratory* studies[5–7], which try to infer new regulatory interactions based on gene expression data. For our initial analysis of the global transcription network of *B. subtilis* (Fig. 2 and Supplementary Fig. 2), we used datasets from both categories. In total, we compared four reconstructions of the global transcription network of *B. subtilis*, assembled in three independent studies (Supplementary Table 1):

(1) Sierro and colleagues (2008)

Sierro and colleagues[1] assembled the Database of Transcription Regulation in *Bacillus subtilis* (DBTBS). This database is no longer updated, but still forms one of the most complete data sets of transcriptional regulation in *B. subtilis*, because for each regulatory interaction included in the database it keeps track of the methods that have been used to document it. We obtained the latest version of this data in October 2016.

(2) Arrieta-Ortiz and colleagues (2015)

Arrietta-Ortiz and colleagues[7] studied the global transcription network of *B. subtilis* with the aim of discovering new regulatory interactions from transcriptomic data. Their study provides two reconstructions of the global transcription network: (1) one reconstruction before training (based on previous literature), used as starting point, and (2) the reconstruction after training. In the training procedure, hundreds of expression profiles were analyzed to infer new regulatory interactions, using a Bayesian inference method that was seeded by the network reconstruction *before* training. Both network reconstructions were included in our analysis for global modularity in transcriptional regulation.

(3) Michna and colleagues (2016)

The SubtiWiki database[4] provides the most comprehensive database on transcriptional regulation in *B. subtilis* and is assembled in the course of approximately 10 years[4,8–11]. The database is still being updated and therefore provides the most up-to-date reconstruction of the global transcription network of *B. subtilis*. We used the network reconstruction available from the SubtiWiki 2.0 database on June 2017.

Even though the above studies characterized the global transcription network independently from each other, they can best be viewed as a cumulative effort to understand transcriptional regulation in *B. subtilis*. Partly for historical reasons, the network structures they infer are similar. For example, the network reconstruction of Arrieta-Ortiz and colleagues[7] before training was partly based on an older version of the network reconstruction available in the SubtiWiki database. For the largest part of our study, we rely on the network reconstruction available on the SubtiWiki website. Supplementary Table 1 gives a full description, including all regulatory interactions, of each of the four network reconstructions discussed in this section.

Supplementary Text 2. Motivation and background of neural network analysis.

High-dimensional gene expression data is typically analyzed using two different approaches. Some studies perform *feature selection*, by focusing on genes of special interest and examining if they are differentially expressed across conditions, whereas others perform *feature extraction*, in which dimensionality reduction techniques are used to extract the general features that best describe a high-dimensional gene expression dataset. We combine both techniques, by extracting genes underlying lifestyle biology in *B. subtilis* from expression data, and using dimensionality reduction to characterize the expression of these genes across a large set of expression profiles.

Standard dimensionality reduction techniques (e.g. Principal Component Analysis) perform linear transformation of data onto a set of orthogonal axes. They work well to segregate gene expressing profiles obtained from a small number of discrete conditions (e.g. low and high temperatures). However, when the number of tested conditions increases, and when these conditions are not discrete (e.g. temperature gradients), expression profiles often show non-linear correlations. In that case, non-linear dimensionality reduction can be superior to capture statistical structure in data. A machine learning method known as autoassociative artificial neural networks (or autoencoders) has been shown to be particularly effective in this type of non-linear dimensionality reduction[12,13]. Besides capturing non-linear correlations, this method has several advantages. First, autoassociative artificial neural networks can be trained to reduce noise. Gene expression data is often inherently noisy, due to the methods of RNA extraction. Such measurement noise has no biological relevance and should be filtered out. When a neural network is trained properly (i.e. no over-training[14]), it will learn how to ignore this type of noise, thereby only extracting biologically relevant information. Second, when modest imbalances exist in the amount of data from different conditions, autoassociative artificial neural networks can compensate for such imbalances and interpolate between sparsely distributed data points[15]. For example, given that

gene expression data is easier to acquire from mid-exponential growth, expression profiles from this growth stage are typically overrepresented in expression databases, whereas profiles from the late stationary growth stage are typically underrepresented. These biases also exist in the dataset of Nicolas and colleagues (2012)[16], where most expression profiles were obtained near the motility peak, whereas relatively few profiles come from the late stages of sporulation (see Fig. 4a and Supplementary Fig. 8). An autoassociative neural network can capture broad trends in the data despite such sampling biases, and capture the entire life cycle of *B. subtilis*, from germination to sporulation (Fig. 3).

The dataset of Nicolas and colleagues (2012)[16] solely contains population-level gene expression data. This has important implications for neural network analysis. Since expression levels are acquired at the population level, they are averages of the expression levels of individual cells. A motility gene could appear to have a low expression level if many cells express the gene weakly, or if few cells express it strongly, whereas others do not express it at all. Based on what we know from microscopy data, such as microfluidic experiments[17,18], cell differentiation in *B. subtilis* is largely stochastic. In other words, in any particular environment, cells have a certain probability to express the genes underlying for example motility or sporulation (sometimes contingent on the past decision of a cell[19]). These probabilities will differ across environments, e.g. cells are more likely to sporulate when they are starving. In consequence, *B. subtilis* populations typically consist of a mixture of different cell types[20–23]. Some cells can be motile, others filamentous, and yet others sporulating. The sequential life stages in Fig. 4a reflect both gene expression changes within cells, as well as changes in the relative abundance of different cell types (see Supplementary Fig. 9). Just after germination, there are mostly motile cells, then more and more cells express genes required for colony formation and, finally, cells start to trigger the endosporulation pathway. The same reasoning holds for the sequential expression of regulators shown in Fig. 4b. Not every cell will express each of the regulators sequentially, but at the population level

these regulators show a regular expression order. Thus, in summary, the gene expression changes summarized in Fig. 4 do not represent homogeneous and deterministic expression changes across all cells in the population, but a statistical map of the average expression changes cells experience as they transition through their distinct life stages in a population (Supplementary Fig. 9).

A neural network analysis is only effective when expression data is dominated by a certain type of directionality or order. That is, life stages and regulators should be expressed in a specific order, such as the transition from motility, to colony formation, to sporulation. If any such order is absent, it becomes impossible to robustly compress data onto a single dimension, because many different dimensions would fit the data equally well. In the case of *B. subtilis*, directionality probably arises from the saprophytic nature of the life cycle. Spores germinate in nutrient-rich conditions. When these resources do not replenish, cells will be confronted with steadily worsening conditions, as they deplete their resources, until cells eventually starve and trigger sporulation. One could say that spores demarcate the life cycle by providing a fixed beginning and ending of this cycle. This does not imply that every individual cell undergoes the same order of phenotypic changes or that changes are irreversible. Rather, it shows that gene expression changes tend to occur in a prevalent order at the population level. This order is captured by the alternation of life stages that make up a life cycle. Genes that are not expressed in any particular life stage, for example because they are triggered by specific conditions only, will not correlate with the pseudotime dimension. Life cycles that are marked by a fixed beginning and ending, such as a dormant life stage, are not restricted to the *Bacilli*, but occur across the phylogenetic tree of the Bacteria (see some examples listed in Supplementary Table 4). Whenever sufficient gene expression data will become available for species in other taxonomic groups, a neural network analysis like the one we applied to *B. subtilis* could be used to study their life cycles, which may uncover many still unexplored life stages and expression modules.

Supplementary Text 3. Regulation of sporulation in a nutshell.

The roles of global regulators in lifestyle switches in *B. subtilis* have been studied intensely, and especially with respect to endosporulation[24–26]. Four sigma factors, namely sigma factors E, F, G and K, play a key role in coordinating the sporulation process[24,27] (Supplementary Fig. 3). In brief, upon starvation, sporulation is triggered by the transcription factor Spo0A, which – when phosphorylated at sufficiently high levels – stimulates the expression of the sigma factors E (σ^E) and F (σ^F)[28–30]. Both sigma factors are inactive until after asymmetrical cell division, which results in the formation of a mother cell and forespore, σ^F becomes active in the forespore, where it controls the expression of tens of genes involved in the further development of the forespore[31] (e.g. engulfment). It induces the expression of *spoIIR* and *sigG*, the gene encoding sigma factor G[32,33], σ^G. SpoIIR crosses the membrane of the forespore and is involved in activating σ^E in the mother cell[33,34]. σ^G is the last sigma factor expressed in the forespore and responsible for completing spore development, such as the production of chromosome condensation proteins and germination receptors[35]. Like for the other sigma factors, σ^G is inactive upon its first expression, but later gets activated by proteins under the control of σ^E in the mother cell[36]. In this way, communication between the mother cell and forespore assures that the developmental process is kept in check[25]. σ^E in the mother cell is involved in numerous sporulation processes as well, such as engulfment of the spore, formation of the spore cortex, and the formation of the spore coat[36]. σ^E also stimulates the expression of the final sigma factor, σ^K, expressed in the mother cell[37]. This sigma factor is first inactive, but then gets activated by SpoIVB, a protein under the control of both σ^F and σ^G in the forespore[38–40]. σ^K is involved in the production of the spore coat as well as the lysis of the mother cell, which releases the spore into the environment[41,42].

The role of global regulators is also well studied for lifestyle changes other than sporulation[2,22,43]: For example, sigma factor D (σ^D) is important for motility, sigma factor H (σ^H) is involved in the transition to

186 stationary growth, ComK is important for competence[20,21,44], and Spo0A is also important for colony

187 formation[45,46].

188

Supplementary Text 4. Phylogenetic tree of the *Bacillales*.

We compared different methods for constructing the phylogenetic tree of the *Bacillales* to make sure that the tree presented in Fig. 5 (Supplementary Fig. 21) is not biased by the employed methodology. We focused on three aspects of our phylogenetic analysis: (1) the set of marker genes that we used to compare the different genomes, (2) the alignment algorithm that we used to align the marker genes, and (3) the trimming algorithm, which we used to trim the aligned marker genes before concatenation.

The criterion used to select the marker genes is likely to have the largest influence on the resulting phylogenetic tree. Ideally, marker genes should be (1) universally present among all genomes in the phylogenetic group of interest, and (2) only occur in single copies in any one genome. These two criteria maximize the chance that gene copies are orthologs and are thus related by descent rather than by duplication or horizontal gene transfer, thereby minimizing incongruencies between the gene tree and the species tree that we seek to reconstruct. Unfortunately, the more species and genomes are considered in a phylogenetic analysis, the fewer genes will meet these stringent criteria. One needs to strike the right balance between applying the above criteria stringently, resulting in few genes suitable for phylogenetic analysis and a potential inability to resolve the tree for closely-related species, and applying the criteria too leniently, yielding many genes for phylogenetic analysis, but possibly obscuring the 'true' phylogenetic signal due to incongruences between gene trees and the species tree.

We start our analysis from a set of marker genes that was previously identified by Wu and colleagues[47,48] for the phylum of the Firmicutes (to which the Bacillales belong). These marker genes were identified based on the strictest criteria, but since fewer genomes were available at the time then today, we re-evaluated these genes and categorized them into three hierarchical groups:

(1) **Marker genes that strictly meet the above criteria**: genes that are universally present in all genomes, with only a single copy per genome (n=4 genes).

(2) **Marker genes that meet intermediate criteria**: genes that occur in single copies in all but four or fewer genomes, where they either have paralogs or are absent altogether (n=53 genes). We note that this set of genes includes the four genes from (1).

(3) **Marker genes that meet weaker criteria**: all marker genes that were previously identified for the Firmicutes, even if absent or having paralogs in 5 or more genomes included in this study (n=87 genes).

To determine the ideal set of marker genes, we constructed phylogenetic trees for each set of marker genes, using six different methods for sequence alignment and trimming. For alignment, we employed the following algorithms: MAFFT v7.299, MUSCLE v3.8.31 and PRANK[49–53]. For trimming, we employed trimAl v1.2 and gBLOCKS v0.91b[54–56]. Thus, in total, we constructed 18 phylogenetic trees (3 sets of marker genes, 3 alignment algorithms and 2 trimming algorithms).

We compared these trees by comparing the average within-tree bootstrap values, as well as the agreement in branching topologies across trees (i.e. branch support). In Supplementary Fig. 22, we show for each tree the average bootstrap values and the average branch support, i.e. the number of trees that have a matching branch, scaled between the lowest and highest values. We applied this scaling in order to emphasize the relative differences between trees. The absolute differences in bootstrap values and branch support were small. As expected, the phylogenetic trees constructed with the smallest set of genes – based on the strictest selection criteria – show the lowest bootstrap values and the lowest branch support. Since this set of genes only includes four highly-conserved genes, it is too small for

resolving the phylogenetic relationship between closely-related strains. In contrast, for genes meeting our intermediate selection criteria, we observed a strong increase in the bootstrap values and branch support (Supplementary Fig. 22). Although there are some differences between trees constructed with the different alignment and trimming methods, these are generally small. For the weakest gene selection criterion, we observed a further slight increase in the average bootstrap values, but this was not associated with a similar increase in the average branch support. In other words, trees from the 'intermediate' and 'weak' sets of genes yield more-or-less the same tree topologies. The increase in bootstrap values might therefore simply result from the larger set of genes in the 'weak' category.

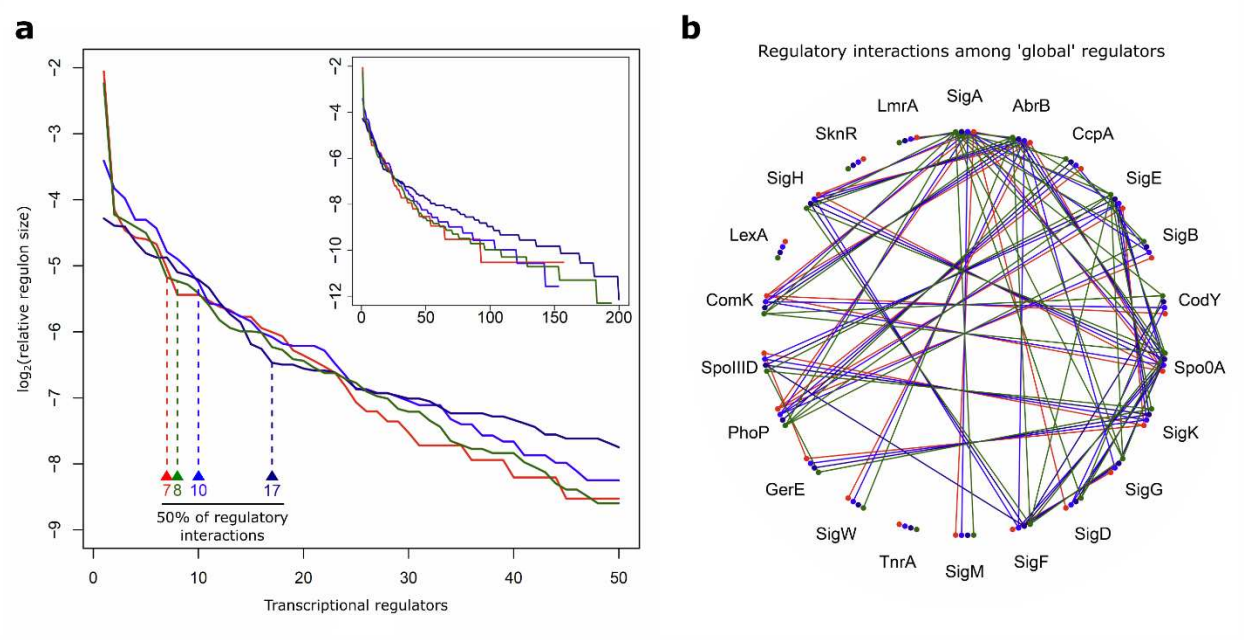
In Supplementary Fig. 23, we explore the general relationship between the within-tree bootstrap values and the across-trees branch support. As expected, branches with high bootstrap support tend to be shared among more trees, even though this pattern is not universal. Some branches have high bootstrap values, but only occur in a small fraction of the trees. This bias might result from incongruences between gene trees, in which a few genes give artificially strong support for a particular branch, whereas the other genes give no or weak support. Excluding these 'misleading' marker genes would immediately change the tree topology. Conversely, other branches have low bootstrap values, but are strongly supported between trees. This bias might result from analyzing too few genes, which may harbor too little phylogenetic information to yield high bootstrap values, even for branches that are generally well supported across trees. To minimize either error, we decided to choose a tree using the 'intermediate' set of genes for our further analysis. This is tree number 8 from Supplementary Fig. 22, which is also the basis of Fig. 5 and Supplementary Fig. 21. We note, however, that similar tree topologies are obtained from any of the other trees constructed from the same set of marker genes. Fig. 5 and Supplementary Fig. 21 visualize bootstrap values of each branch with the size of the circles displayed at the branch: *small* circles indicate low bootstrap values and *large* circles show high bootstrap values. Branch color

indicates how well a branch is supported across trees: branches that are *red* are not shared with other trees and branches that are *green* are shared with all other trees. In Fig. 5 and Supplementary Fig. 21 we collapsed all branches between genomes with a phylogenetic distance smaller than 0.01. The phylogenetic relation between such closely-related genomes could not be resolved based on our set of marker genes. Importantly, we did not collapse those branches for our analyses in Supplementary Figs. 22 and 23, which explains the relatively high fraction of low bootstrap values in these figures compared to the branches that are shown in Fig. 5 and Supplementary Fig. 21.

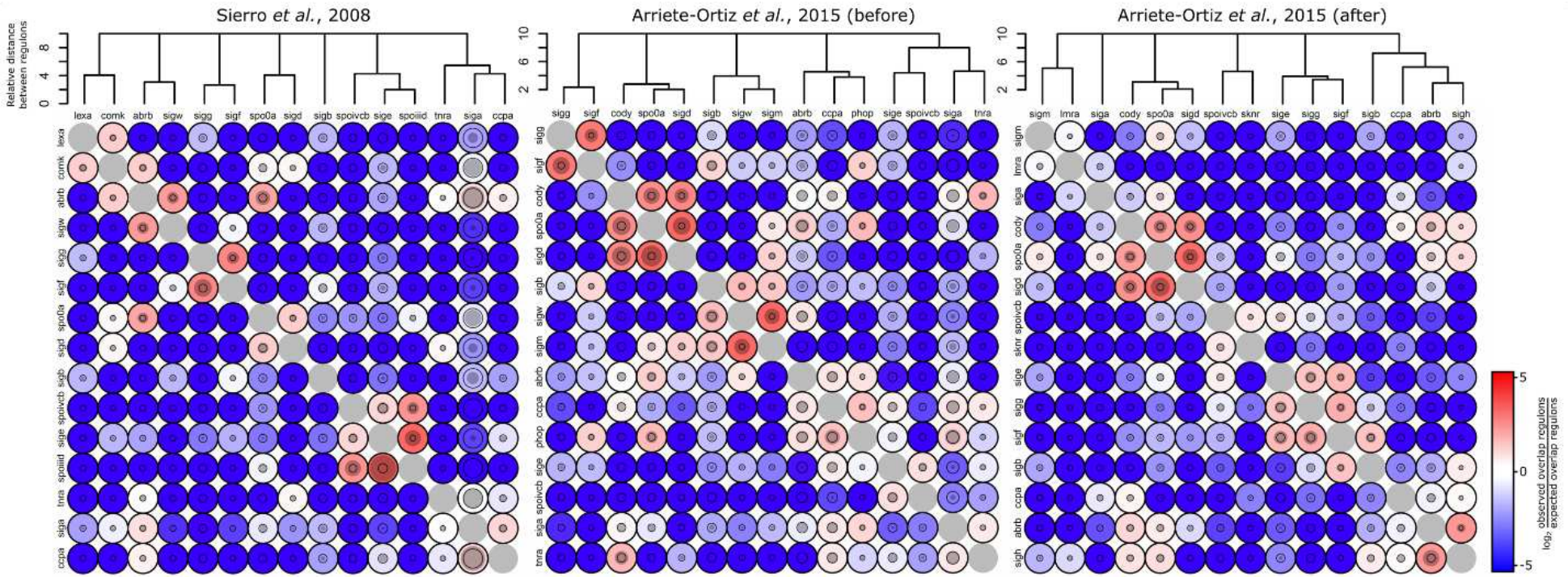
In Fig. 5 we explore how life stages are conserved across the *Bacillales*, by specifically focusing on the regulatory targets of 'life cycle' regulators identified in Fig. 4b. We also explore the conservation of regulators themselves. In general, one expects that regulators with many downstream regulatory targets (e.g. global regulators) are more strongly conserved than regulators with few targets, since mutations in those regulators would have many more pleiotropic effects[57]. This is indeed the case: Supplementary Fig. 24 shows that regulators with large regulons are more-strongly conserved across the *Bacillales* than regulators with small regulons. Global regulators show especially strong conservation. These results hold when determining conservation both in terms of orthology, i.e. bi-directional best BLAST hits, and in terms of unidirectional protein BLAST hits. The same overall pattern is apparent when examining the conservation of genes in the gene regulatory networks underlying lifestyle transitions (Supplementary Fig. 25): regulatory genes are more strongly conserved than either sensory or effector genes. However, for sensory genes, results are strongly dependent on the method by which conservation is determined: when using unidirectional BLAST hits, instead of bi-directional best BLAST hits, conservation levels appear much higher. This indicates that even though species may share few orthologous genes, they do often have paralogous genes resulting from gene duplications or horizontal gene transfer. Since paralogous genes can encode for proteins with redundant functions, it might well

be that – despite a lack of orthologs – sensory functions are strongly conserved across species in the *Bacillales*. To ensure that our conclusions based on the comparative genomic results in Fig. 5 also hold when paralogous genes are included in the analysis, we redid the entire analysis using unidirectional BLAST hits, instead of bi-directional best BLAST hits. Supplementary Fig. 26 shows that life stage losses are also apparent when we include paralogous genes to the analysis.

Supplementary Figures



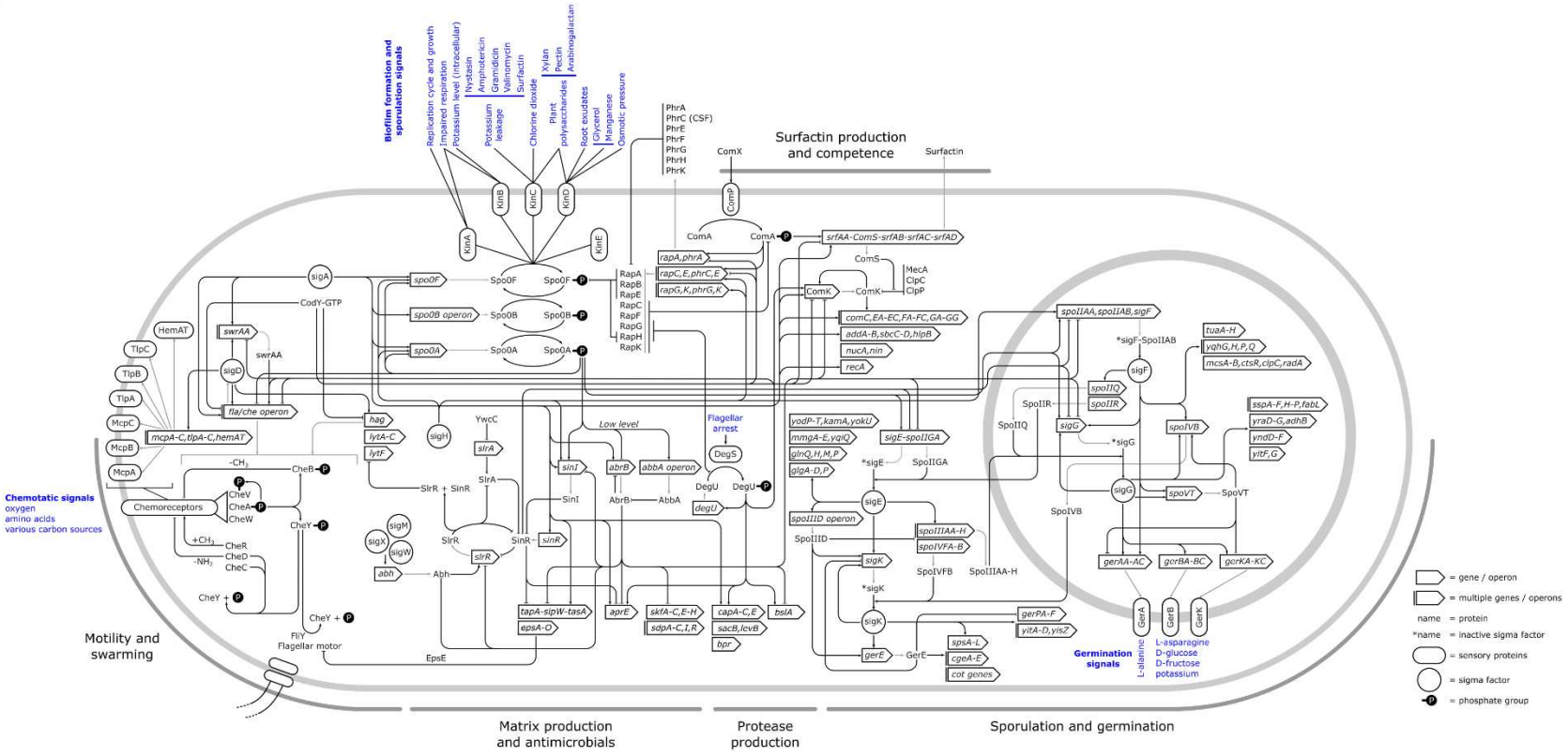
Supplementary Fig. 1. Regulon sizes and interaction between global regulators. a, Transcription regulators sorted by regulon size: red, database of transcriptional regulation in *B. subtilis* (DBTBS, Sierro and colleagues[1]); light and dark blue, observed and inferred transcription network of *B. subtilis* by Arrietta-Ortiz and colleagues[7]; green, transcription network of *B. subtilis* based on SubtiWiki database[4]. **b**, Regulatory interactions between the 15 transcriptional regulators with largest regulon sizes. Since the top-15 global regulators are not the same for all databases, a total of 22 regulators is shown in the circular diagram.



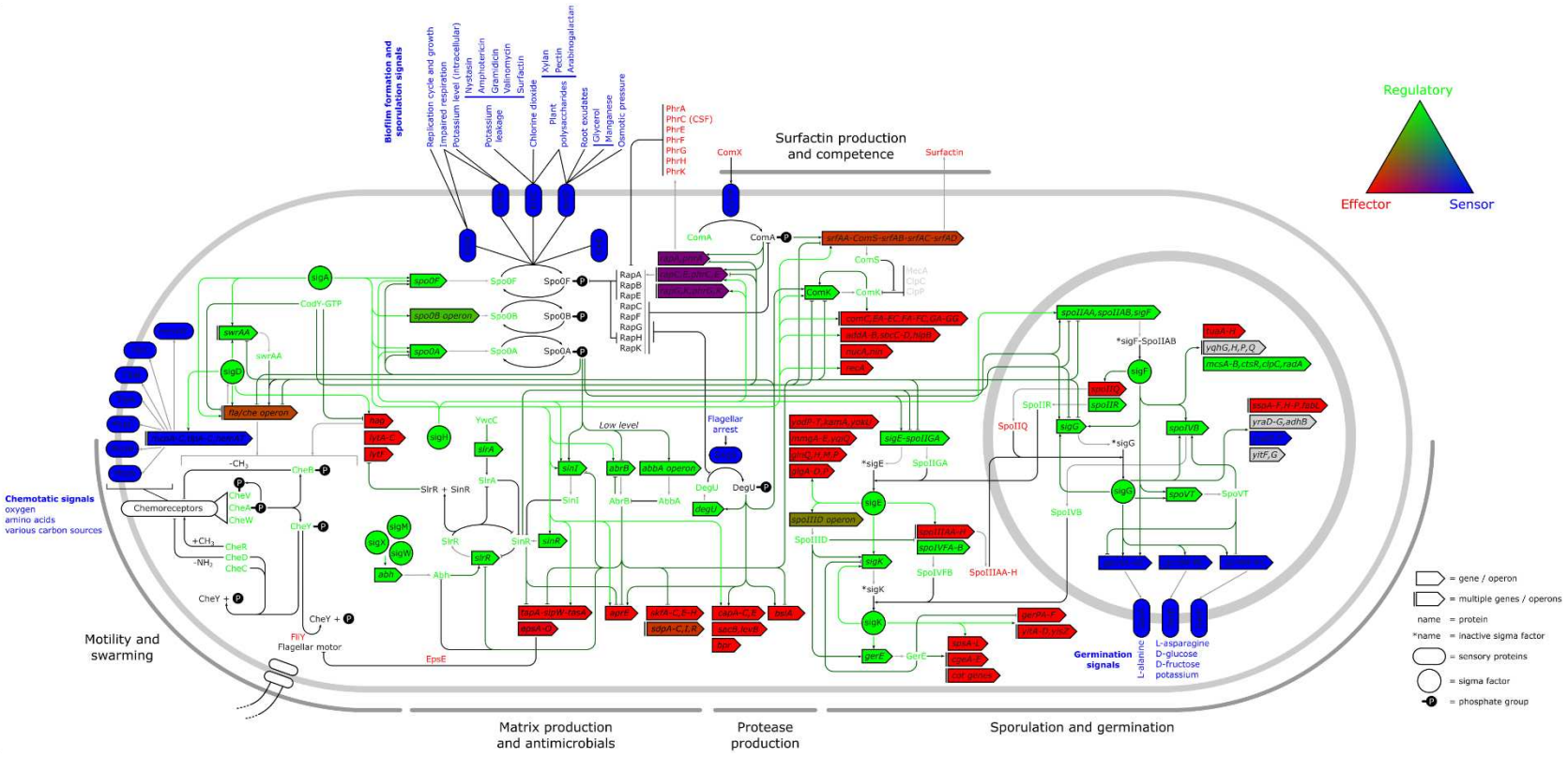
297

298 **Supplementary Fig. 2. Global modularity in transcriptional regulation in other reconstructions.** Same analysis as shown in Fig. 2, but applied to

299 the other reconstructions on the global transcription network of *B. subtilis* (see also Supplementary Text 1)**.**



Supplementary Fig. 3. Overview of gene regulatory network underlying lifestyle switches in *Bacillus subtilis*. Grey arrows, transcription and translation; black arrows, transcriptional or post-transcriptional regulation; blue names, putative environmental and cytoplasmic cues sensed by sensory proteins. Genes belonging to the same operon and operons that have the same upstream regulatory proteins are collapsed into a single component in the diagram (see legend). The network comprises 7% of all genes in *B. subtilis*, covering 107 operons and including 12% of all described regulatory interactions in the global transcription network. These genes and interactions form the regulatory backbone for lifestyle switches in *B. subtilis*.



307

308 **Supplementary Fig. 4. Classification of genes/proteins in gene regulatory network**. Sensory genes encode proteins (blue) that receive signals

309 directly from the internal or external environment and not from an upstream protein. Regulatory genes encode proteins (green) that receive

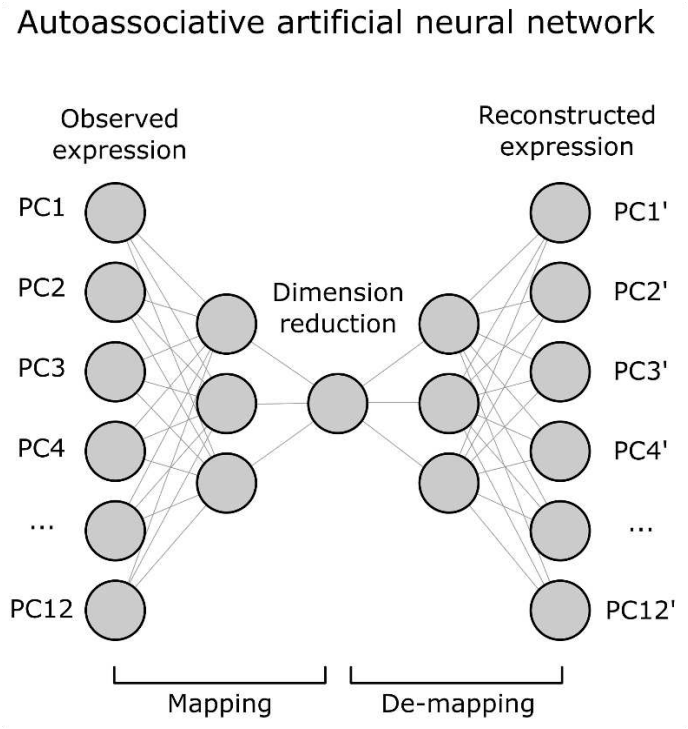
310 signal from upstream signaling/regulatory protein and transfer this information to downstream protein by (post)-transcriptional regulation.

311 Effector genes encode proteins (red) that fulfill any non-regulatory function. Some proteins carry multiple promiscuous functions, making

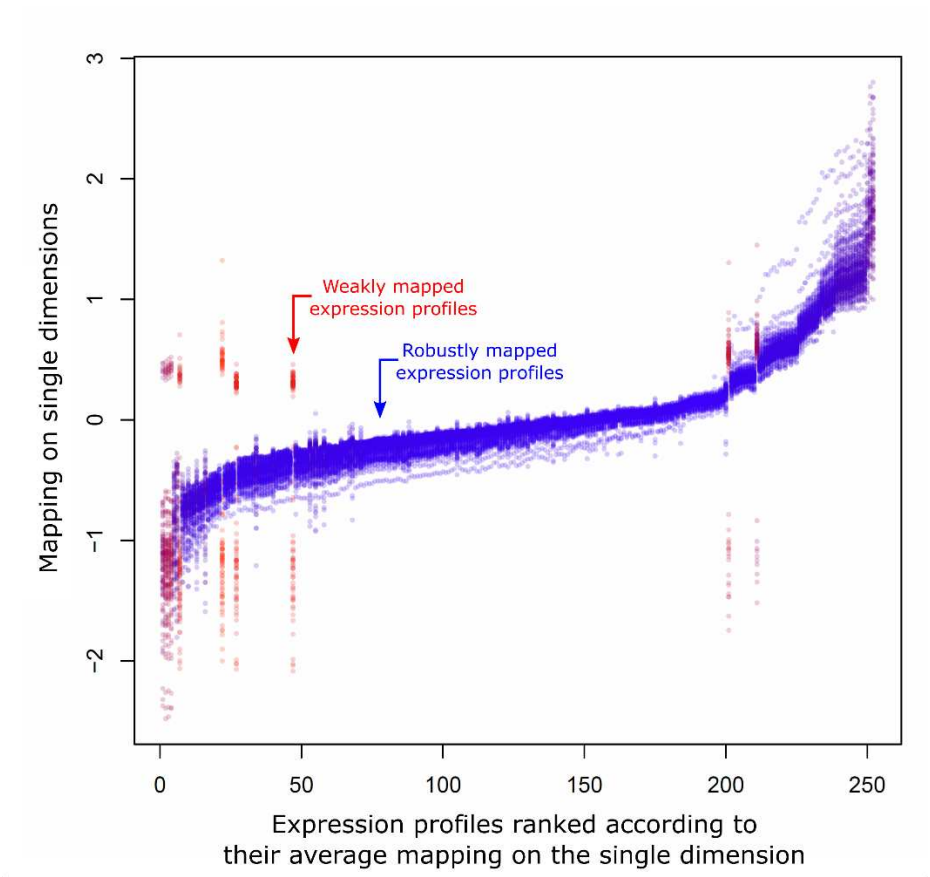
312 categorization difficult (e.g. ComX, which is secreted into the environment, thereby forming an effector protein, but is at the same time sensed

313 by the membrane-bound kinase ComP). Genes and proteins are colored in the same way, but operons that encode multiple genes are colored

according to the RGB-color mapping shown in the legend, based on the relative fraction of sensory, regulatory and effector genes. Genes for which functionality is ambiguous are shown in grey. Green arrows, transcriptional regulation; black arrows, post-transcriptional regulation; grey arrows, transcription and translation.

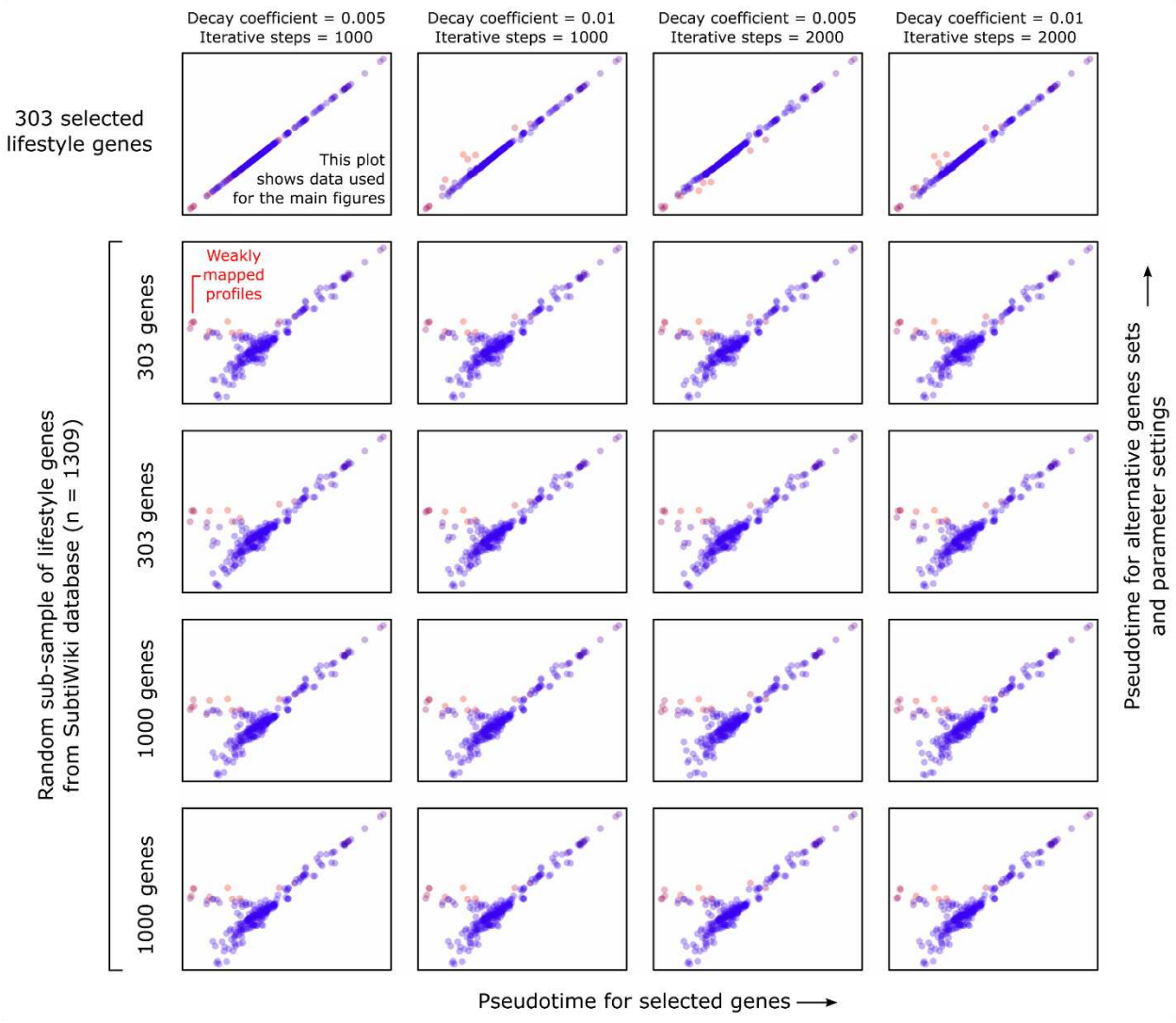


Supplementary Fig. 5. Autoassociative artificial neural network. This neural network is used for non-linear dimensionality reduction of the transcriptional dataset of Nicolas and colleagues[16]. As a first step, the multidimensional gene expression dataset is subjected to a principal component analysis, from which the 12 first components are used as entries to the neural network. The network is trained such that the network entries match the network outputs as closely as possible (see Methods).



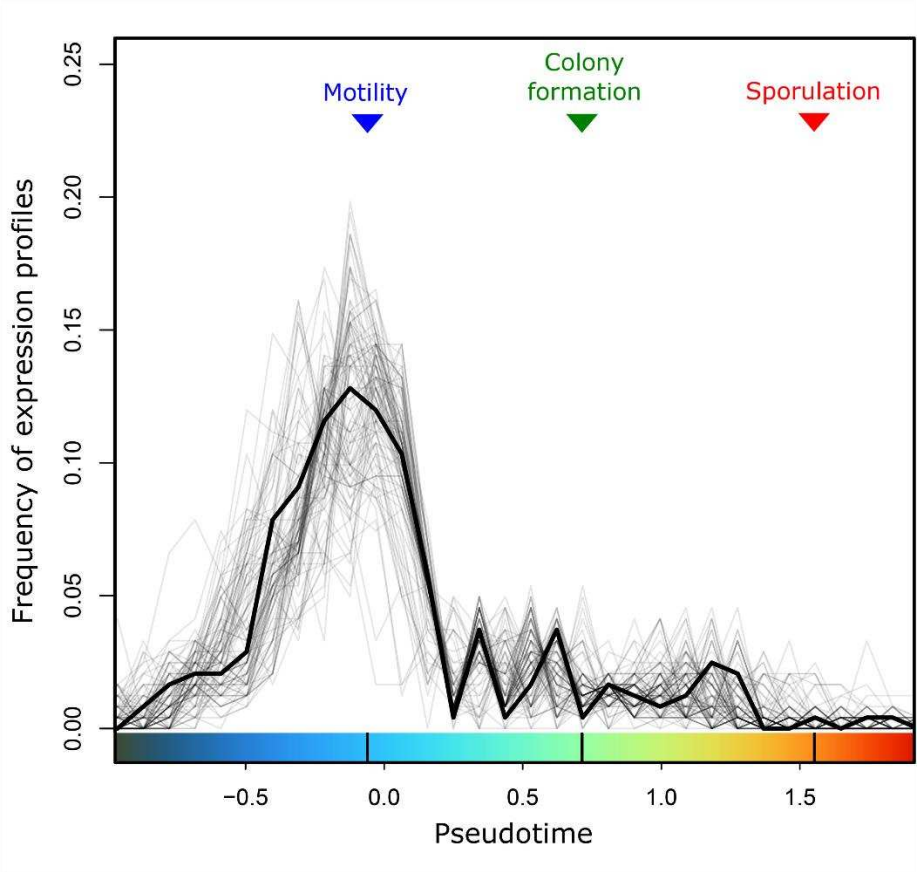
326

327 **Supplementary Fig. 6. Mapping of expression profiles on hundred single dimensions**. Expression
328 profiles are ranked according to their average mapping on the hundred single dimensions, generated by
329 training a hundred independent autoassociative artificial neural network on the same gene expression
330 dataset. For each expression profile, the color shows the relative variation in this profile's mapping on
331 the hundred single dimensions. Profiles with smallest variation are shown in blue and profiles with
332 largest variation are shown in red. Whereas most profiles robustly map to the single dimensions (blue),
333 10 profiles express a weak mapping (red).

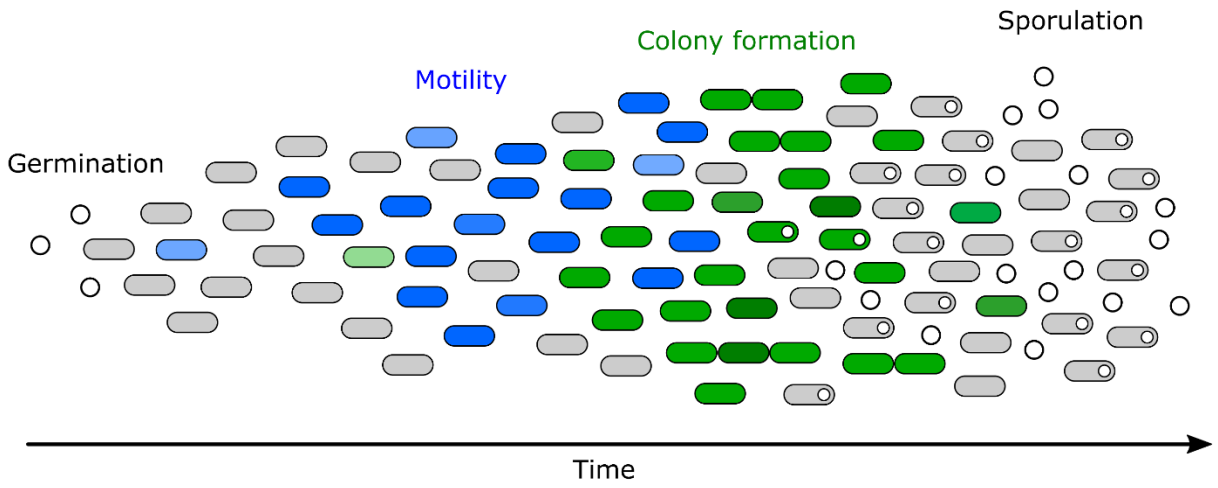


Supplementary Fig. 7. Robustness analysis of autoassociative artificial neural network training. Each plot shows the correlation between the average mapping of expression profiles in Fig. 3b (i.e. pseudotime) and the average mapping of expression profiles for an alternative selection of lifestyle genes and/or an alternative parameter setting in neural network training. Plots along the x-axis of the figure differ in the parameter settings. Specifically, we varied the decay coefficient and number of iterative training steps used to train the neural networks. Plots along the y-axis of the figure differ in the selection of lifestyle genes. Top row corresponds to the selection of 303 lifestyle genes in Supplementary Table 3 and the four lower rows correspond to different random subsamples of lifestyle

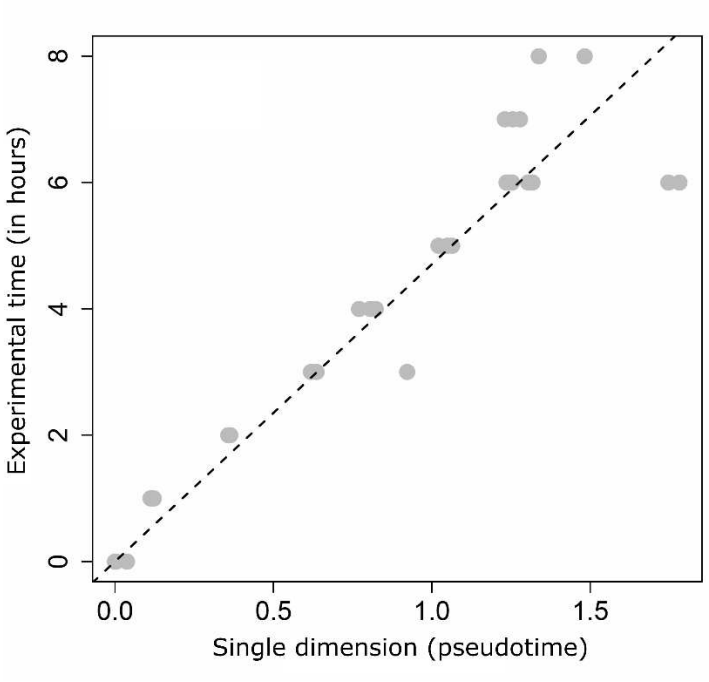
343 genes from the SubtiWiki database[4]. We either randomly selected 303 genes or 1000 genes from the
344 category of lifestyle genes identified in the SubtiWiki database (n = 1309 genes). In all cases, there was a
345 strong correlation between the average mapping of profiles observed in Fig. 3 and the average mapping
346 of profiles for the alternative conditions tested in this figure (for all plots, $P < 10^{-16}$, $R^2 > 0.7$). The colors
347 correspond to those shown in Supplementary Fig. 6. Expression profiles that were weakly mapped in
348 Supplementary Fig. 6 are also more variable in this figure.



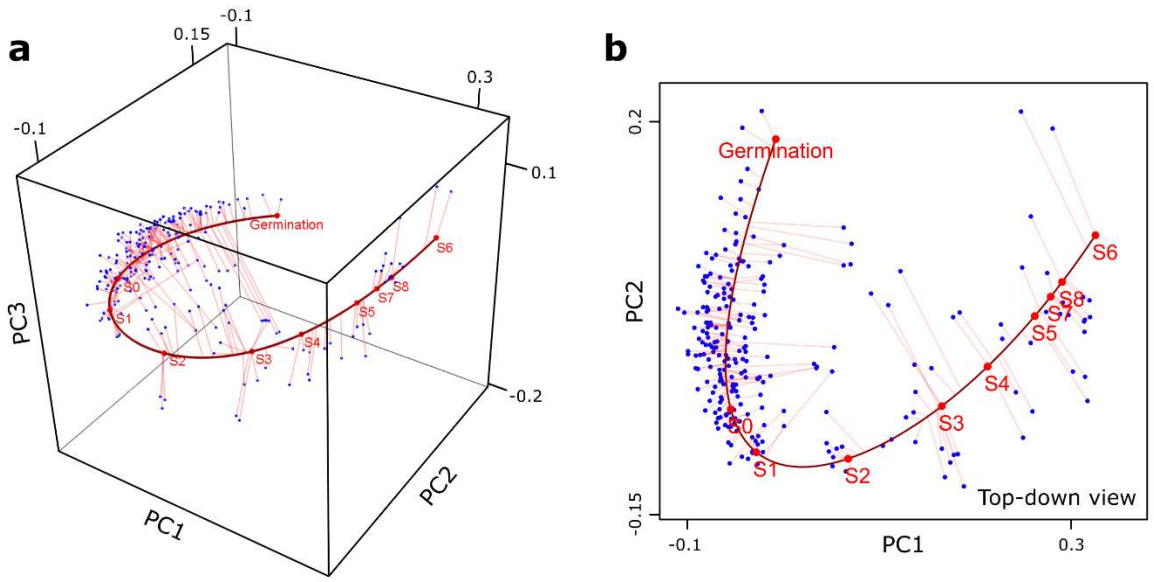
Supplementary Fig. 8. Distribution of expression profiles along pseudotime dimension as shown in Fig 4a. Bold black line, average distribution of expression profiles. Light grey lines, distribution according to 100 independently trained neural networks (Fig. 3b).



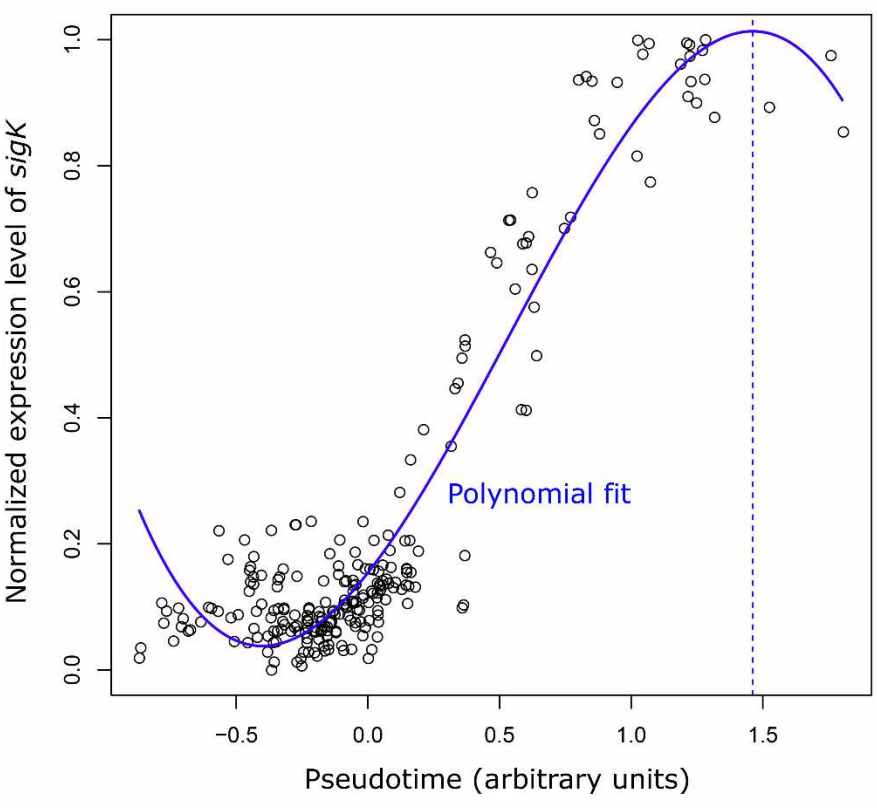
Supplementary Fig. 9. Schematic depiction of life stage transitions captured in life cycle reconstruction. Expression changes reflect both regulatory changes that happen within cells, as well as changes in relative cell type abundance.



Supplementary Fig. 10. Correlation between mapping of expression profiles on single dimension (i.e. pseudotime) and experimental time. The dataset of Nicolas and colleagues[16] includes time series data, in which expression profiles were acquired every hour for the duration of the experiment (8h). This graph shows the correlation between the experimental time at which these expression profiles were acquired and their average mapping (i.e. pseudotime) on the single dimension (dashed line shows linear regression, $R^2 = 0.97$, $P < 10^{-16}$).



Supplementary Fig. 11. Experimental time projected on single dimension. **a**, 3D view and, **b**, top-down view of single dimension produced by training a single autoassociative artificial neural network (corresponding to Fig. 3a). Germination, mapping of expression profile that is acquired at the moment of germination. S0-S8, mapping of expression profiles acquired 0 till 8 hours after the induction of sporulation.



374

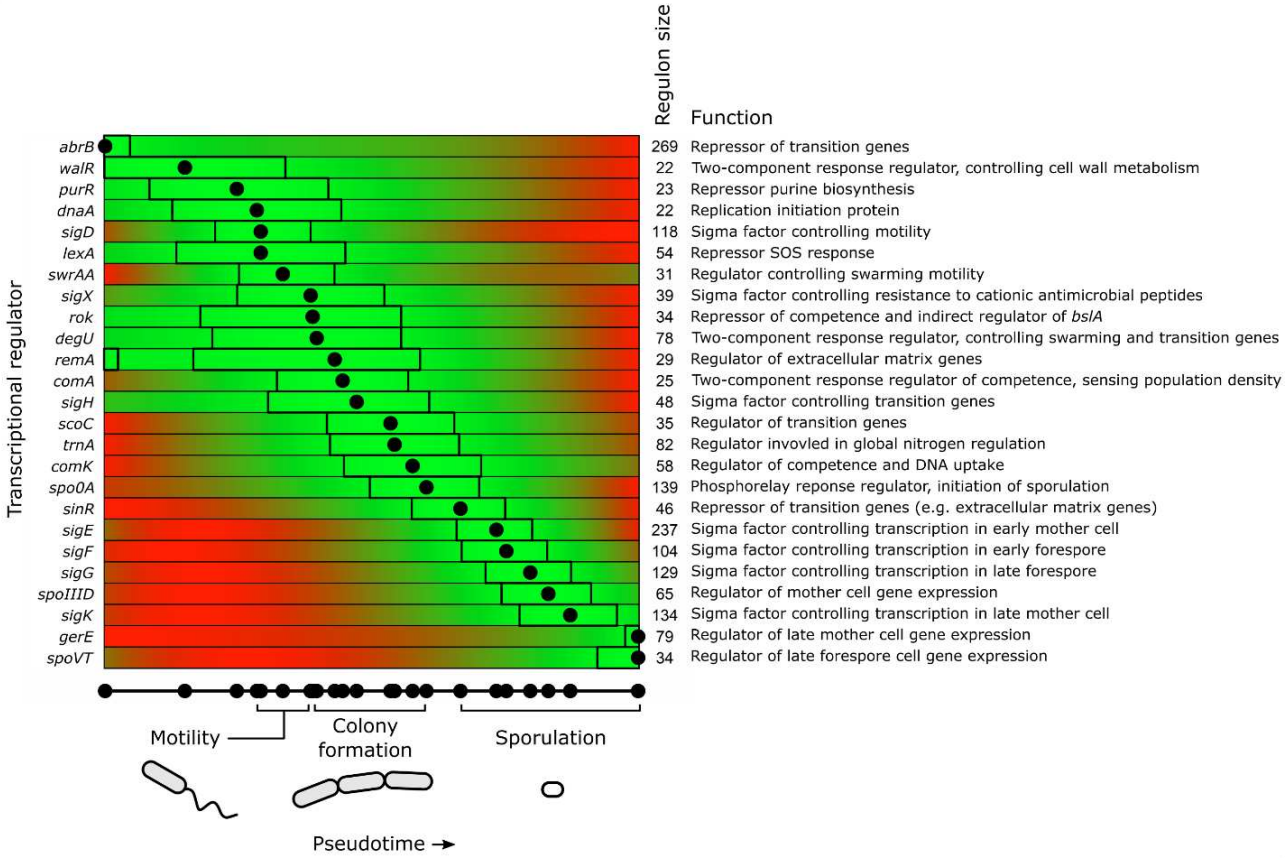
375 **Supplementary Fig. 12. Example of gene expression change in pseudotime of global regulator *sigK*.**

376 *sigK* encodes sigma factor K, involved in late sporulation development. Blue solid line, polynomial

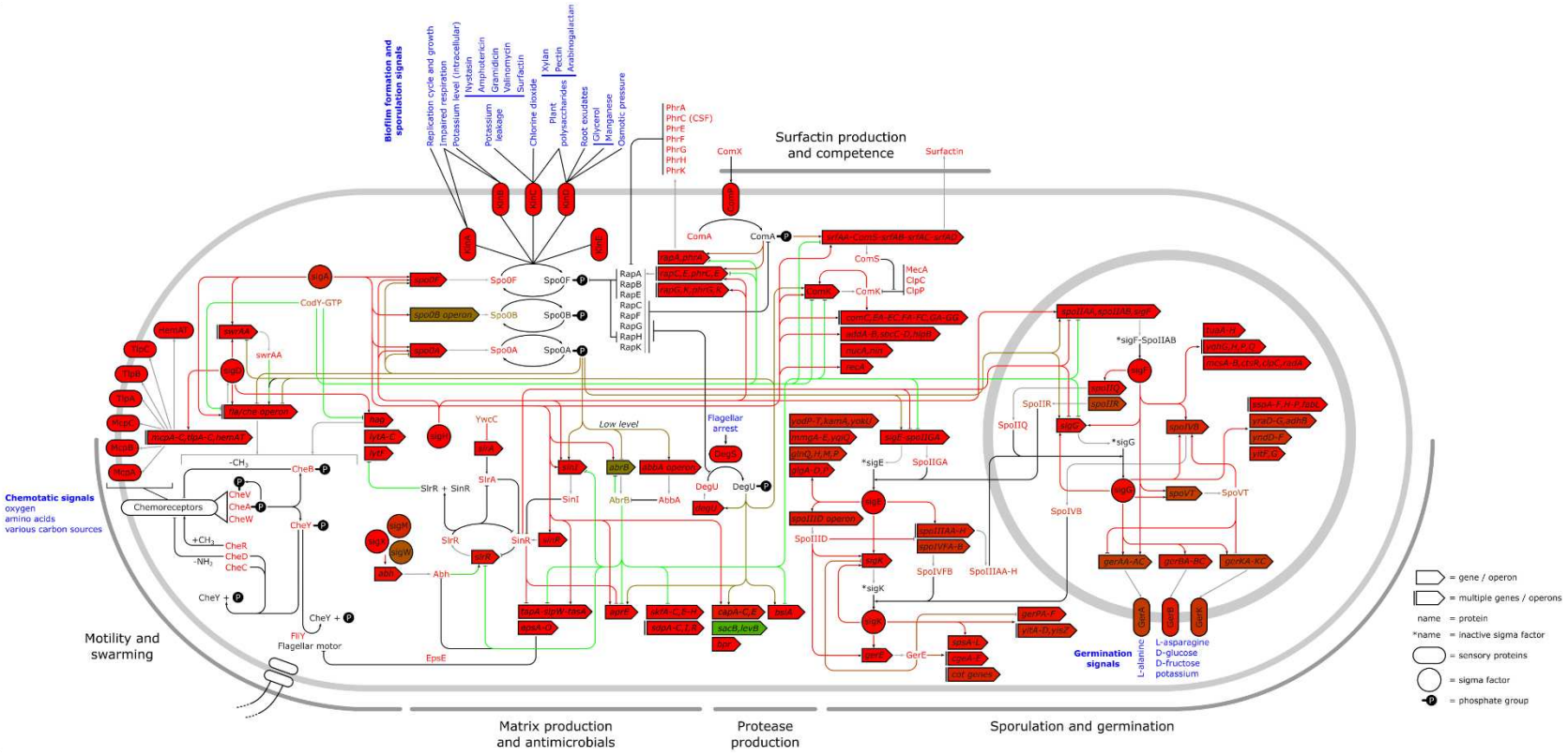
377 regression ($R^2 = 0.89$, $P < 10^{-5}$). Blue dashed line, expression peak according to polynomial regression.

378 This peak corresponds to the expression peak shown in Fig. 4b,c.

379



380 **Supplementary Fig. 13. Expression of regulators in pseudotime.** This diagram shows a more detailed
381 depiction of Fig. 4b. The horizontal bars show the polynomial regressions of gene expression in
382 pseudotime for each of the regulators, normalized between the lowest (red) and highest (green)
383 expression levels. The boxes show the highest 5% expression levels and the dots correspond to the
384 expression peak (the grey bar and black dot in Fig. 4b, respectively). The expression peaks are also
385 mapped upon pseudotime axis at the bottom of the graph, with the corresponding life stages. For each
386 regulator, the regulon size and function, according to the SubtiWiki database[4], is shown on the right.



387

388 **Supplementary Figs. 14. Expression levels and protein activity for time point 1 highlighted in Fig. 4b**. Gene expression: red/green genes show

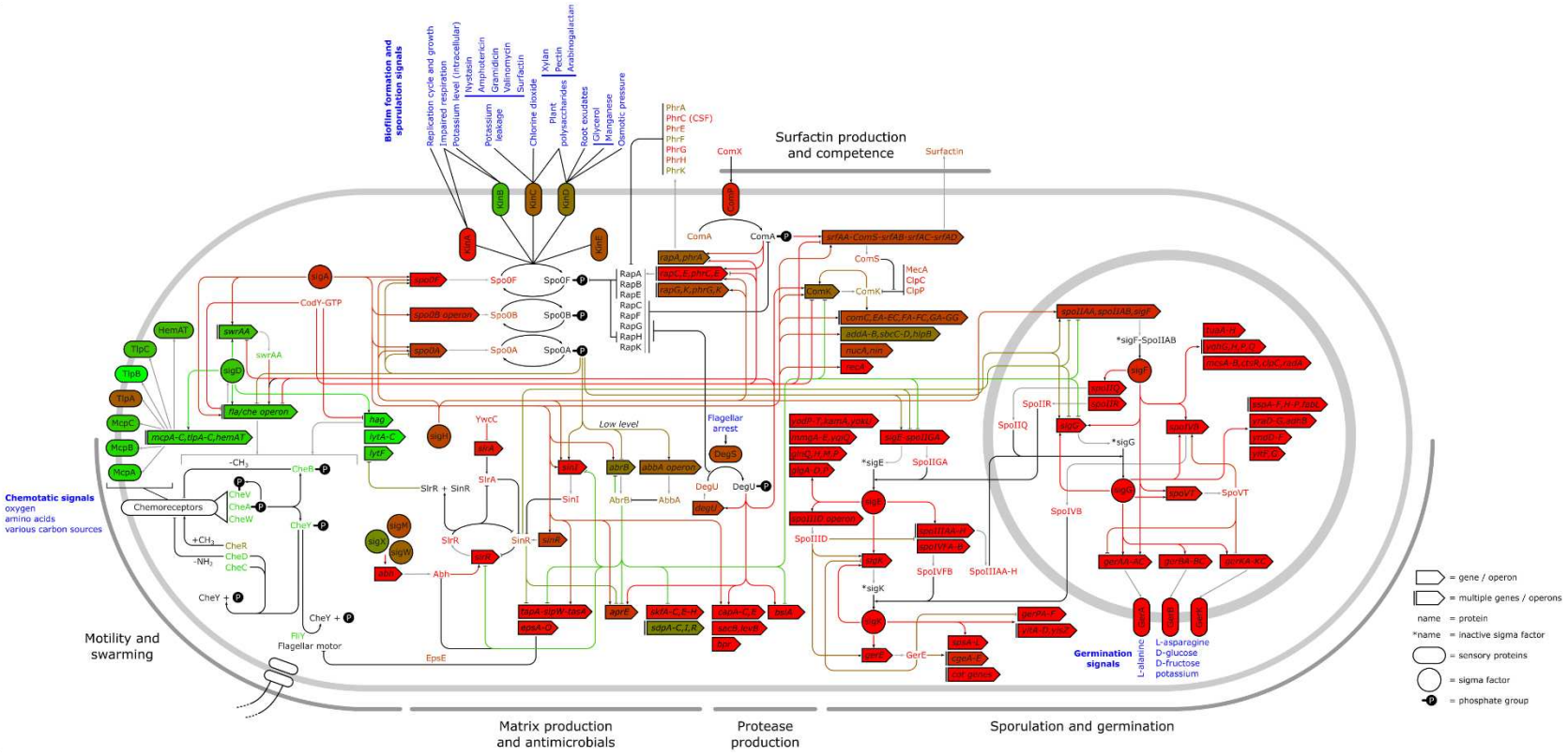
389 low/high expression levels. For simplicity, genes and proteins are colored in the same way. Protein activity is determined in accordance with

390 Arrieta-Ortiz and colleagues[7]: red/green lines show low/high protein activity. Line colors only represent the strength of activity, not the logic

391 (e.g. repression or induction). Expression levels or activity values are normalized between the 10th (red) and 90th (green) percentile, to minimize

392 the effect of outliers on the linear color mapping.

393



394

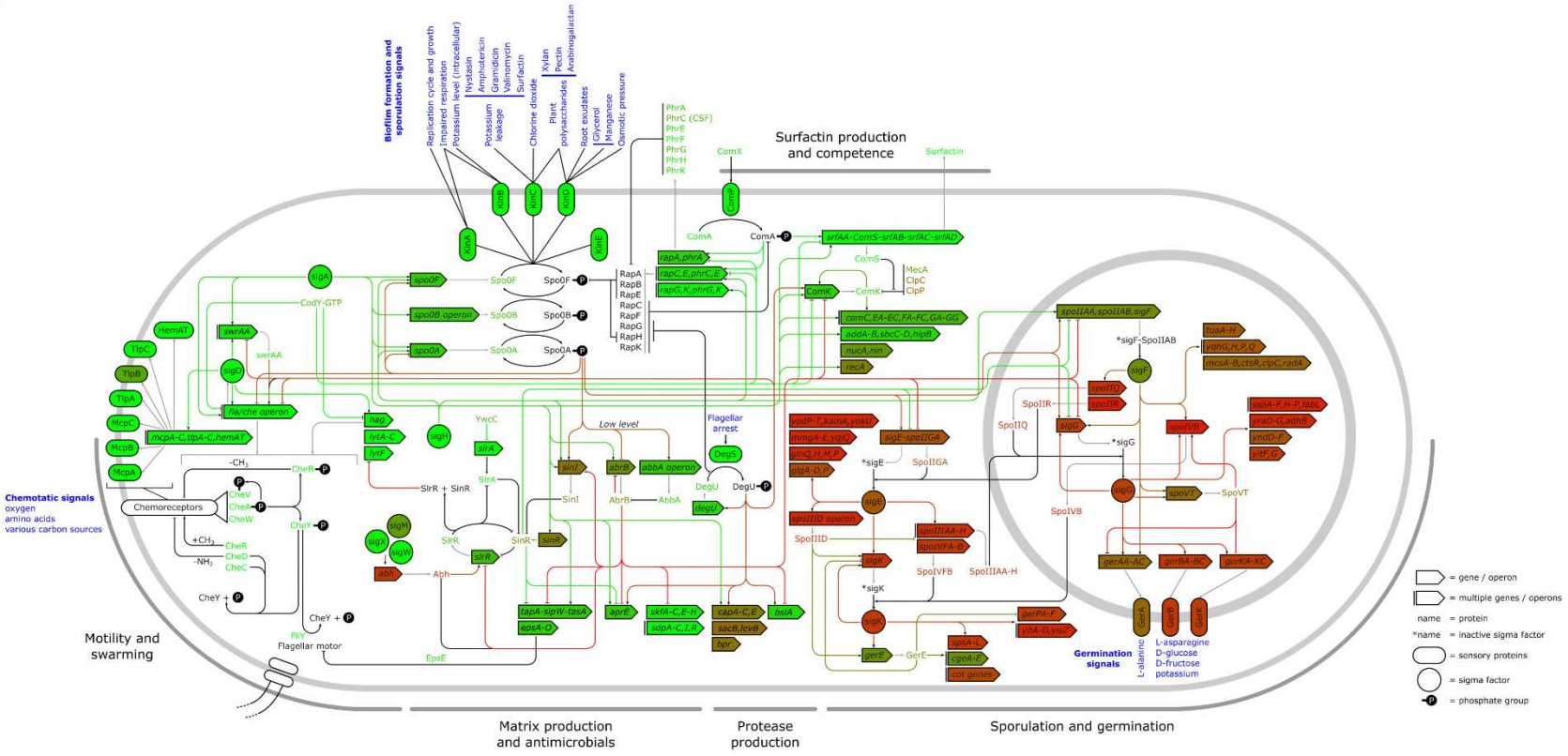
395 **Supplementary Fig. 15. Expression levels and protein activity for time point 2 highlighted in Fig. 4b** Gene expression: red/green genes show

396 low/high expression levels. For simplicity, genes and proteins are colored in the same way. Protein activity is determined in accordance with

397 Arrieta-Ortiz and colleagues[7]: red/green lines show low/high protein activity. Line colors only represent the strength of activity, not the logic

398 (e.g. repression or induction). Expression levels or activity values are normalized between the 10th (red) and 90th (green) percentile, to minimize

399 the effect of outliers on the linear color mapping.



400

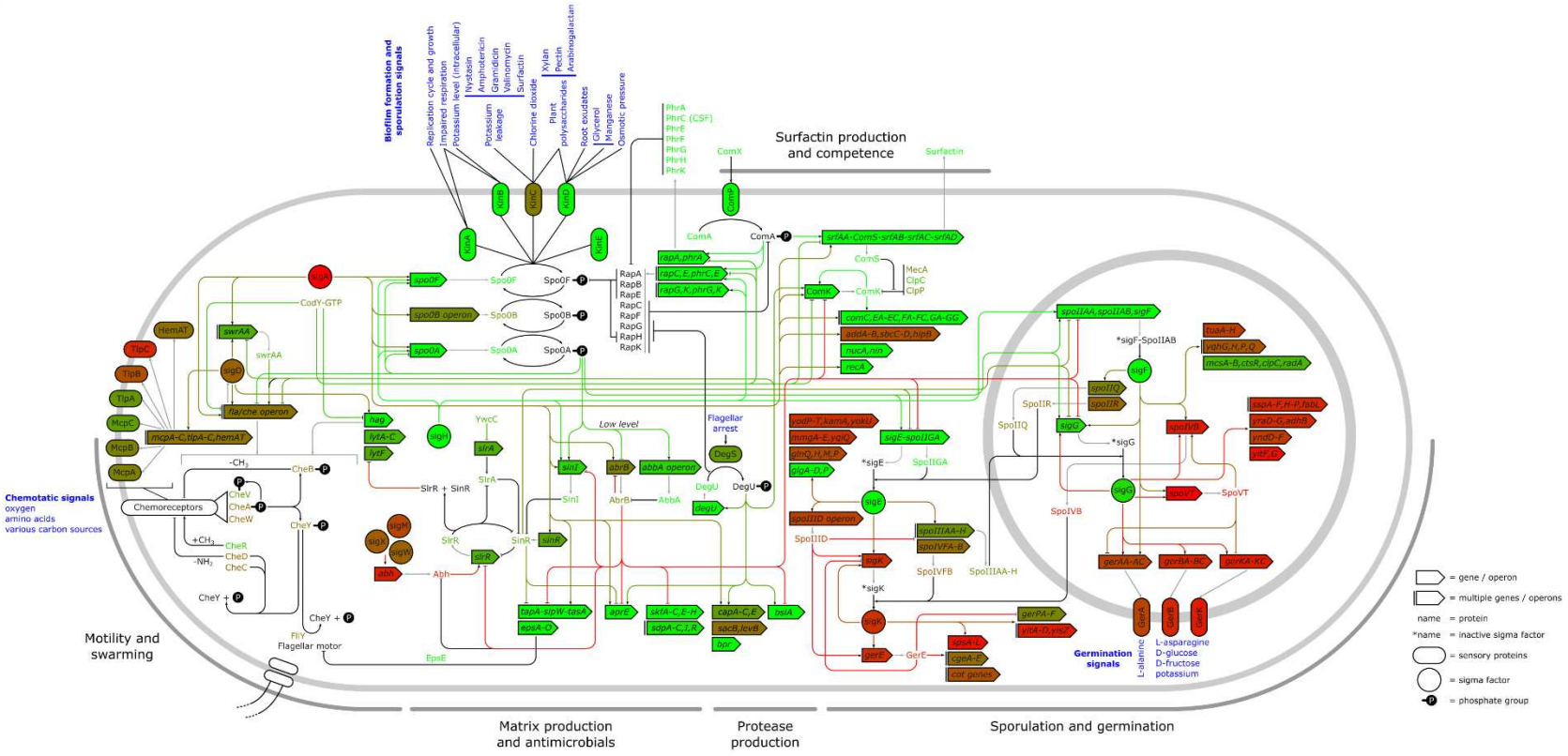
401 **Supplementary Fig. 16. Expression levels and protein activity for time point 3 highlighted in Fig. 4b** Gene expression: red/green genes show

402 low/high expression levels. For simplicity, genes and proteins are colored in the same way. Protein activity is determined in accordance with

403 Arrieta-Ortiz and colleagues[7]: red/green lines show low/high protein activity. Line colors only represent the strength of activity, not the logic

404 (e.g. repression or induction). Expression levels or activity values are normalized between the 10th (red) and 90th (green) percentile, to minimize

405 the effect of outliers on the linear color mapping.



406

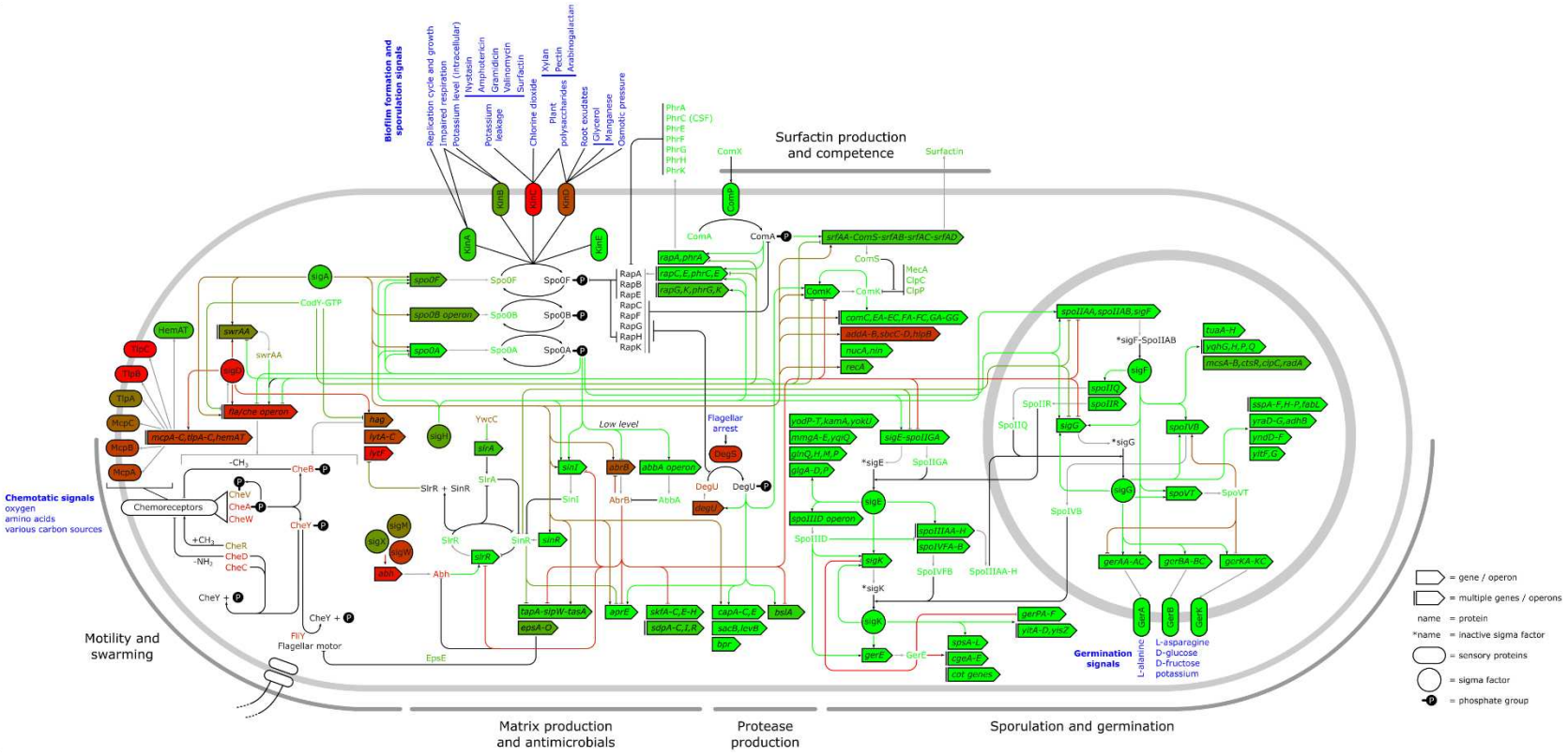
407 **Supplementary Fig. 17. Expression levels and protein activity for time point 4 highlighted in Fig. 4b** Gene expression: red/green genes show

408 low/high expression levels. For simplicity, genes and proteins are colored in the same way. Protein activity is determined in accordance with

409 Arrieta-Ortiz and colleagues[7]: red/green lines show low/high protein activity. Line colors only represent the strength of activity, not the logic

410 (e.g. repression or induction). Expression levels or activity values are normalized between the 10th (red) and 90th (green) percentile, to minimize

411 the effect of outliers on the linear color mapping.



412

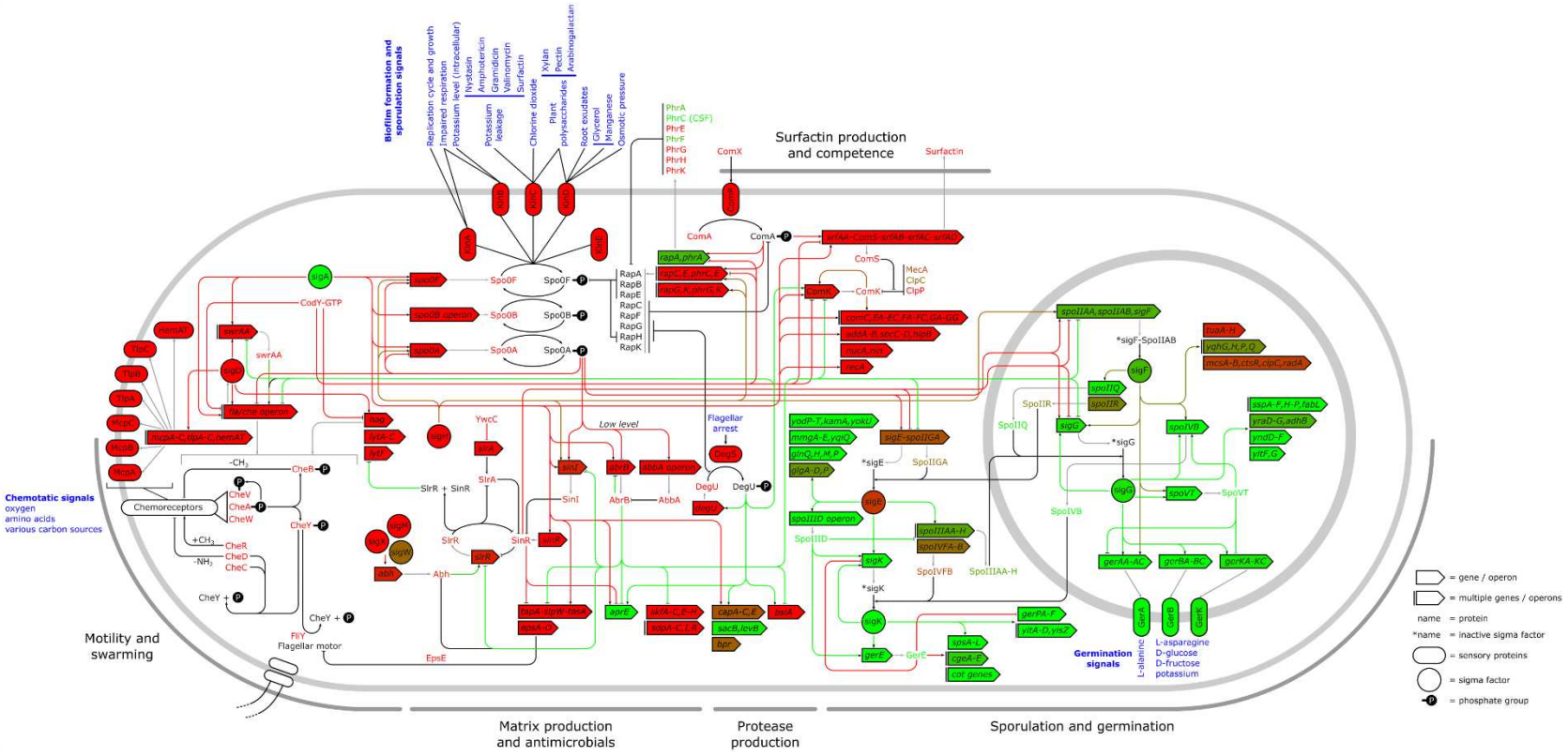
413 **Supplementary Fig. 18. Expression levels and protein activity for time point 5 highlighted in Fig. 4b** Gene expression: red/green genes show

414 low/high expression levels. For simplicity, genes and proteins are colored in the same way. Protein activity is determined in accordance with

415 Arrieta-Ortiz and colleagues[7]: red/green lines show low/high protein activity. Line colors only represent the strength of activity, not the logic

416 (e.g. repression or induction). Expression levels or activity values are normalized between the 10th (red) and 90th (green) percentile, to minimize

417 the effect of outliers on the linear color mapping.



418

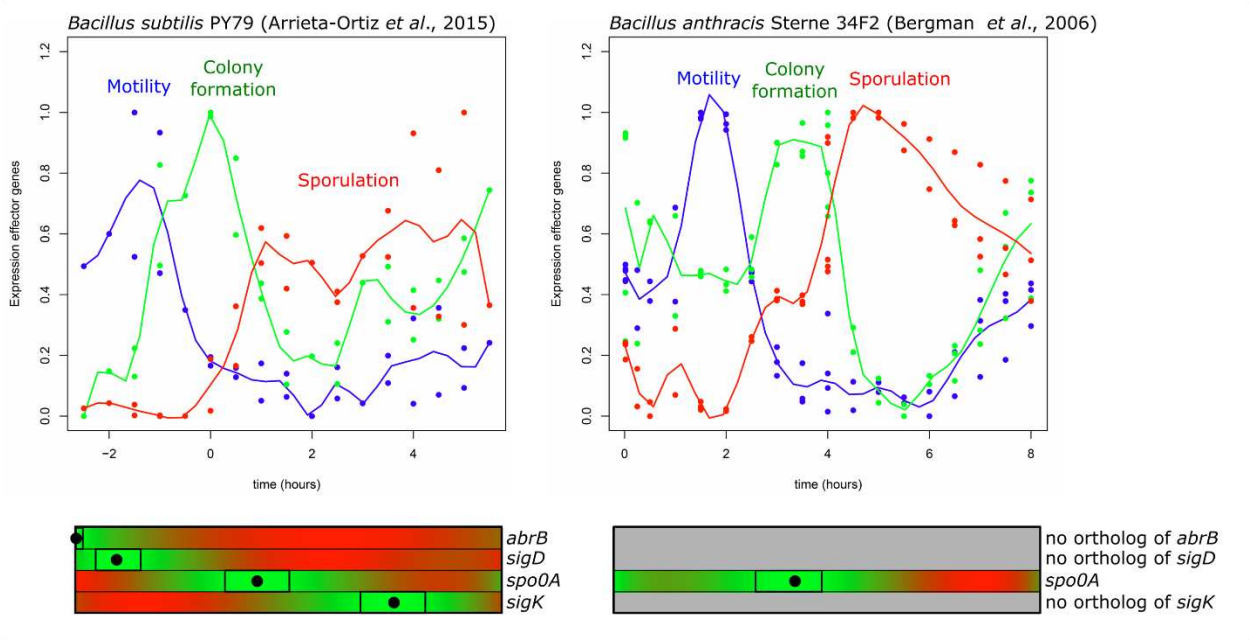
419 **Supplementary Fig. 19. Expression levels and protein activity for time point 6 highlighted in Fig. 4b.** Gene expression: red/green genes show

420 low/high expression levels. For simplicity, genes and proteins are colored in the same way. Protein activity is determined in accordance with

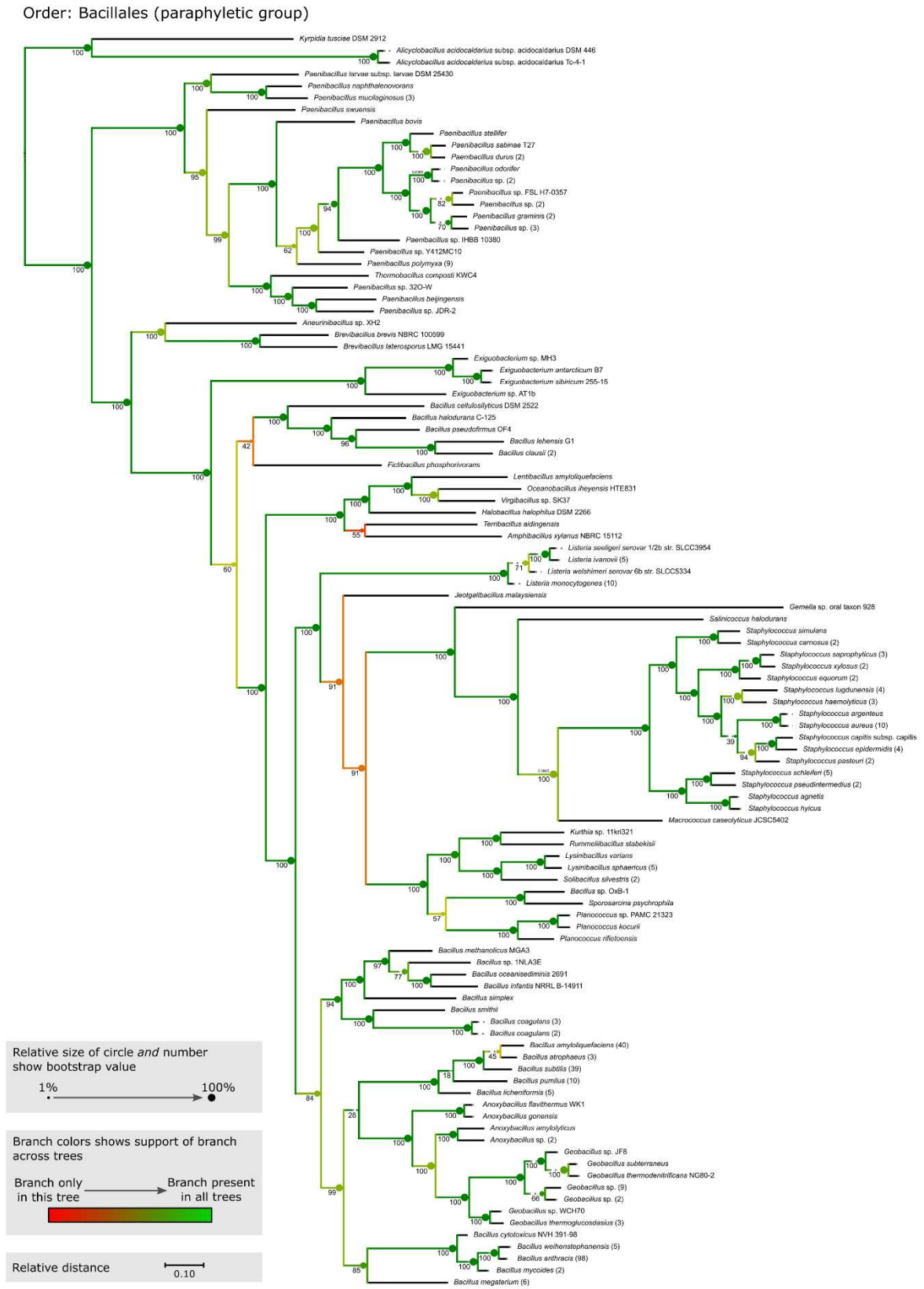
421 Arrieta-Ortiz and colleagues[7]: red/green lines show low/high protein activity. Line colors only represent the strength of activity, not the logic

422 (e.g. repression or induction). Expression levels or activity values are normalized between the 10th (red) and 90th (green) percentile, to minimize

423 the effect of outliers on the linear color mapping.

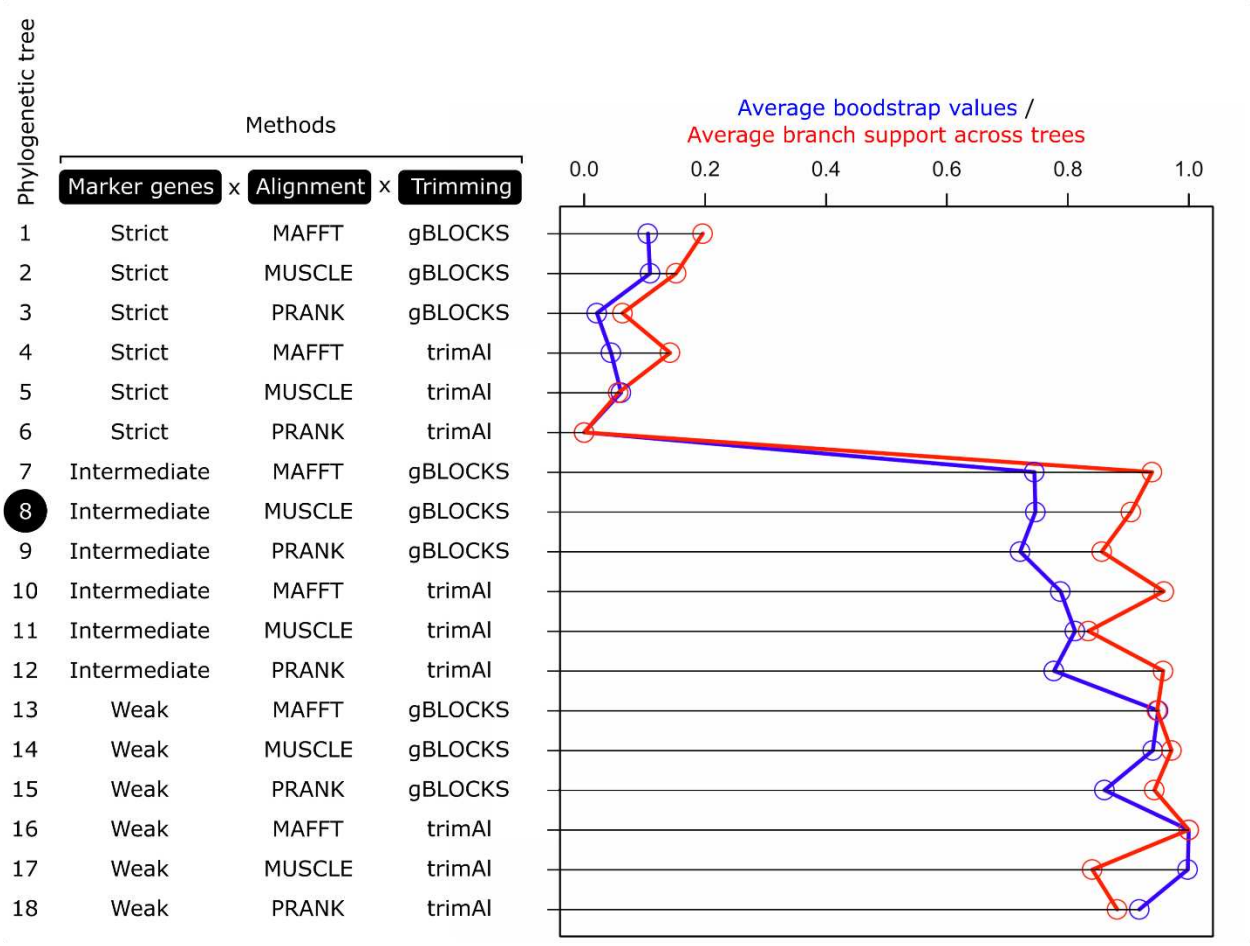


Supplementary Fig. 20. Expression of lifestyles in time-series experiment of related *Bacillales* strain and species. Expression of lifestyle genes in transcriptomic time series data of *Bacillus subtilis* PY79 and *Bacillus anthracis* Sterne 34F2[7,58]. The lifestyle genes in *B. subtilis* PY79 and *B. anthracis* Sterne 34F2 are selected by determining gene orthology to the lifestyle genes in Supplementary Table 3 (see Methods). As apparent in our life cycle reconstruction for *B. subtilis* 168, three clearly distinct life stages can be identified in *B. subtilis* PY79 and *B. anthracis* Sterne 34F2: motility (blue), colony formation (green) and sporulation (red). Expression is normalized between 0 and 1 and solid lines show spline interpolation (n=30). At the bottom of the graphs the polynomial regressions are highlighted for four regulators, in accordance with Supplementary Fig. 13. In *Bacillus anthracis* Sterne 34F2, there is only one orthologous gene of these regulators, as determined by the bi-direction BLAST hits. This does not exclude the possibility that paralogous regulators are present.

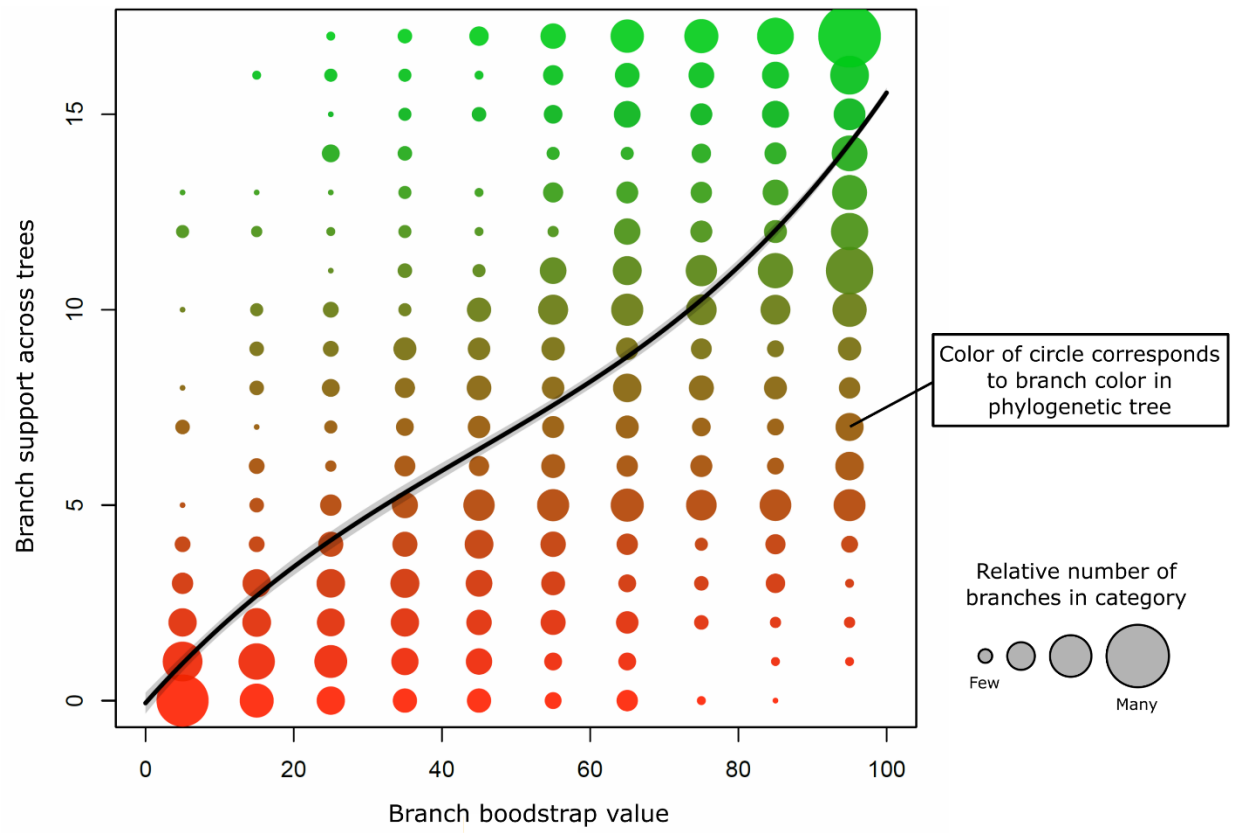

Order: Bacillales (paraphyletic group)
Kyrpidia tusciae DSM 2912
Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446
Alicyclobacillus acidocaldarius subsp. acidocaldarius Tc-4-1
Paenibacillus larvae subsp. larvae DSM 25430
Paenibacillus naphthalenovorans
Paenibacillus mucilaginosus (3)
Paenibacillus swuensis
Paenibacillus bovis
Paenibacillus stellifer
Paenibacillus sabinae T27
Paenibacillus durus (2)
Paenibacillus odorifer
Paenibacillus sp. (2)
Paenibacillus sp. FSL H7-0357
Paenibacillus sp. (2)
Paenibacillus graminis (2)
Paenibacillus sp. (3)
Paenibacillus sp. IHBB 10380
Paenibacillus sp. Y412MC10
Paenibacillus polymyxa (9)
Thermobacillus composti KWC4
Paenibacillus sp. 32O-W
Paenibacillus beijingensis
Paenibacillus sp. JDR-2
Aneurinibacillus sp. XH2
Brevibacillus brevis NBRC 100599
Brevibacillus laterosporus LMG 15441
Exiguobacterium sp. MH3
Exiguobacterium antarcticum B7
Exiguobacterium sibiricum 255-15
Exiguobacterium sp. AT1b
Bacillus cellulosilyticus DSM 2522
Bacillus halodurans C-125
Bacillus pseudofirmus OF4
Bacillus lehensis G1
Bacillus clausii (2)
Fictibacillus phosphorivorans
Lentibacillus amyloliquefaciens
Oceanobacillus iheyensis HTE831
Virgibacillus sp. SK37
Halobacillus halophilus DSM 2266
Terribacillus aidingensis
Amphibacillus xylanus NBRC 15112
Listeria seeligeri serovar 1/2b str. SLCC3954
Listeria ivanovii (5)
Listeria welshimeri serovar 6b str. SLCC5334
Listeria monocytogenes (10)
Jeotgalibacillus malaysiensis
Gemella sp. oral taxon 928
Salinicoccus halodurans
Staphylococcus simulans
Staphylococcus carnosus (2)
Staphylococcus saprophyticus (3)
Staphylococcus xylosus (2)
Staphylococcus equorum (2)
Staphylococcus lugdunensis (4)
Staphylococcus haemolyticus (3)
Staphylococcus argenteus
Staphylococcus aureus (10)
Staphylococcus capitis subsp. capitis
Staphylococcus epidermidis (4)
Staphylococcus pasteuri (2)
Staphylococcus schleiferi (5)
Staphylococcus pseudintermedius (2)
Staphylococcus agnetis
Staphylococcus hyicus
Macrococcus caseolyticus JCSC5402
Kurthia sp. 11kri321
Rummeliibacillus stabekisii
Lysinibacillus varians
Lysinibacillus sphaericus (5)
Solibacillus silvestris (2)
Bacillus sp. OxB-1
Sporosarcina psychrophila
Planococcus sp. PAMC 21323
Planococcus kocurii
Planococcus rifietoensis
Bacillus methanolicus MGA3
Bacillus sp. 1NLA3E
Bacillus oceanisediminis 2691
Bacillus infantis NRRL B-14911
Bacillus simplex
Bacillus smithii
Bacillus coagulans (3)
Bacillus coagulans (2)
Bacillus amyloliquefaciens (40)
Bacillus atrophaeus (3)
Bacillus subtilis (39)
Bacillus pumilus (10)
Bacillus licheniformis (5)
Anoxybacillus flavithermus WK1
Anoxybacillus gonensis
Anoxybacillus amylolyticus
Anoxybacillus sp. (2)
Geobacillus sp. JF8
Geobacillus subterraneus
Geobacillus thermodenitrificans NG80-2
Geobacillus sp. (9)
Geobacillus sp. (2)
Geobacillus sp. WCH70
Geobacillus thermoglucosidasius (3)
Bacillus cytotoxicus NVH 391-98
Bacillus weihenstephanensis (5)
Bacillus anthracis (98)
Bacillus mycoides (2)
Bacillus megaterium (6)
Relative size of circle and number show bootstrap value
1%
100%
Branch colors shows support of branch across trees
Branch only in this tree
Branch present in all trees
Relative distance
0.10

436

Supplementary Fig. 21. Phylogenetic tree of the *Bacillales* corresponding to Fig. 5. Bootstrap values are shown by circles and numbers: small circles indicate low bootstrap values and large circles indicate high bootstrap values. Branch support across independently constructed phylogenetic trees are shown by color: red line colors indicate low branch support and green line colors indicate high branch support. Branch support is here defined as the fraction of phylogenetic trees that have a matching phylogenetic branch (see Supplementary Text 4 for details).



Supplementary Fig. 22. Comparison between methods of phylogenetic tree construction. For each of the 18 independently constructed phylogenetic trees, we show the average bootstrap values (blue) and branch support values across trees (red), normalized between the lowest and highest observed values. We purposely show relative differences, as opposed to the small absolute differences, to emphasize the differences between trees. We tested three sets of marker genes, using conservative, intermediate and lenient selection criteria; three types of alignment algorithms, MAFFT v7.299, MUSCLE v3.8.31 and PRANK; and two trimming algorithms, gBLOCKS v0.91b and trimAl v1.2. See Methods and Supplementary Text 4 for details. The phylogenetic tree shown in Fig. 5 corresponds to eighth phylogenetic tree in this figure.



455

456 **Supplementary Fig. 23. Relation between within-tree bootstrap values and across-trees branch**

457 **support**. For each branch, we determined the within-tree bootstrap value and the across-trees branch

458 support. Since we constructed 18 phylogenetic trees (see Supplementary Fig. 22), the branch support

459 could have a maximum value of 17, corresponding to a branch in a specific focal tree that is shared

460 among all other trees. The bootstrap values are binned in steps of 10%. The plot shows the positive

461 relation between the bootstrap values and branch support (line and shade show polynomial regression

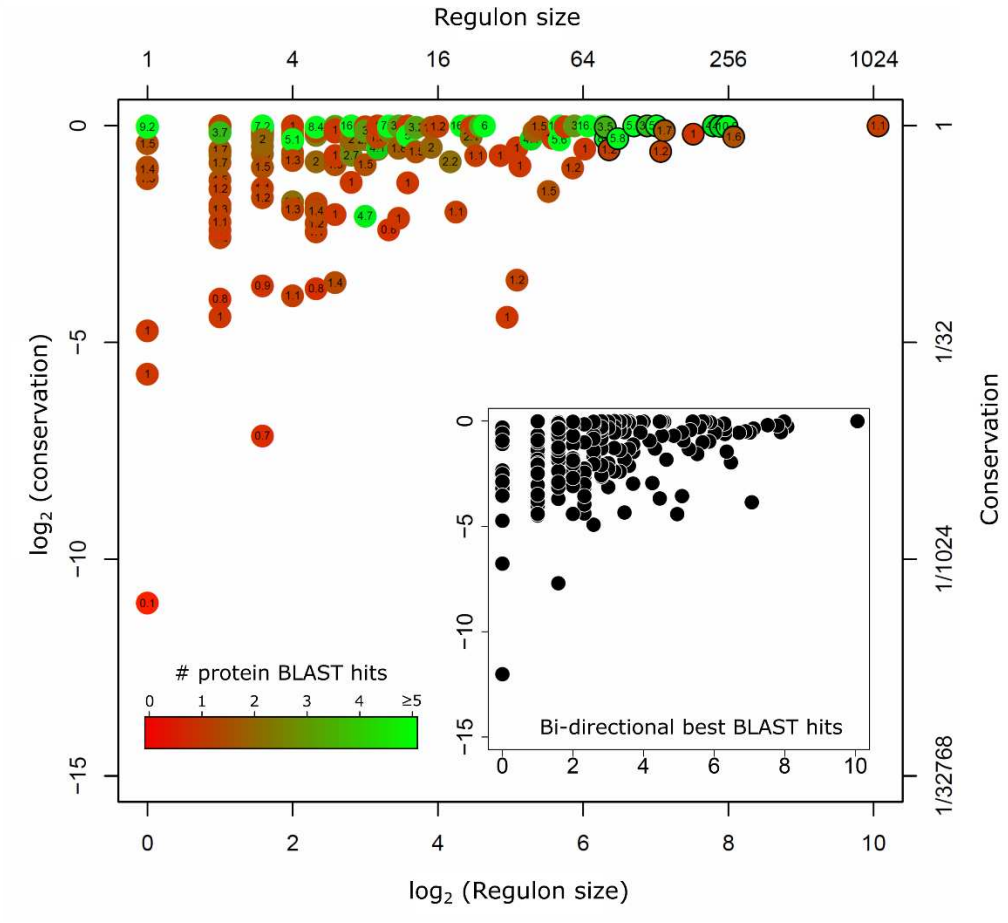
462 and 95% confidence intervals; $R^2 = 0.7$, $P < 10^{-15}$). The relatively high fraction of low bootstrap values is

463 explained by the high number of closely-related genomes that are included in our study. For Fig. 5 and

464 Supplementary Fig. 21, all branches with a phylogenetic distance smaller than 0.01 are collapsed (see

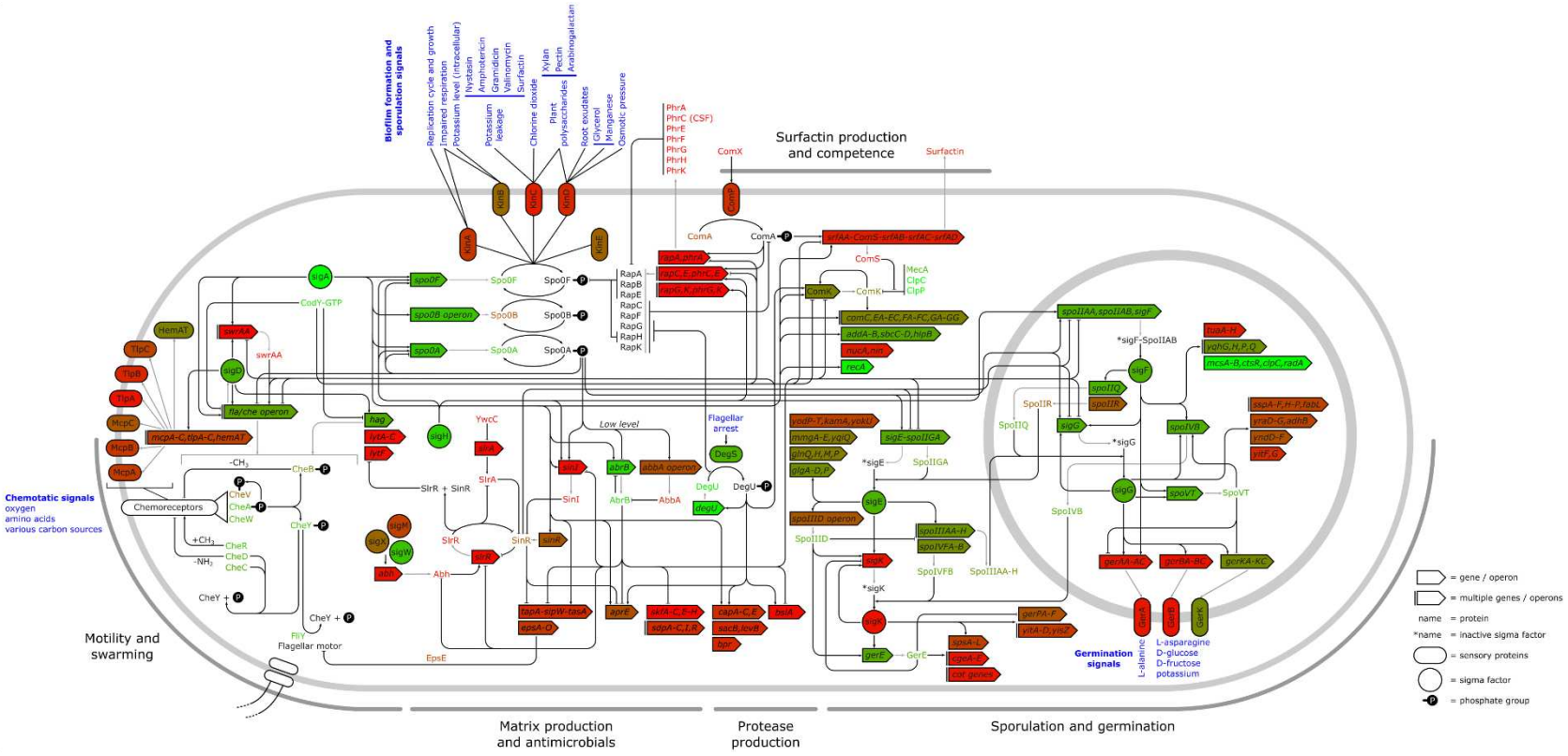
465 Methods), thereby grouping the closely-related genomes for which the phylogenetic relationship could

466 not be resolved.



467

468 **Supplementary Fig. 24. Conservation of regulators based on regulon size.** Conservation of regulators in
469 the global transcription network of *B. subtilis* in relation to regulon size. Conservation is here
470 determined by average fraction of genomes with at least one protein BLAST hit (e-value < 10^{-10} and
471 coverage > 70%) across all branches of the phylogenetic tree (Supplementary Fig. 21, see also
472 Supplementary Text 4). The numbers and colors show the average number of protein BLAST hits within
473 branches with at least one positive hit: red = 0 average hits and green ≥ 5 average hits. The average
474 number of hits can be lower than 1, in case branches are associated with multiple genomes, of which
475 only a fraction has a positive hit. The global regulators (defined as the 15 regulators with the largest
476 regulon sizes) are shown by the black outlined points. The insert shows the same results for bi-
477 directional best BLAST hits as a conservation criterion.



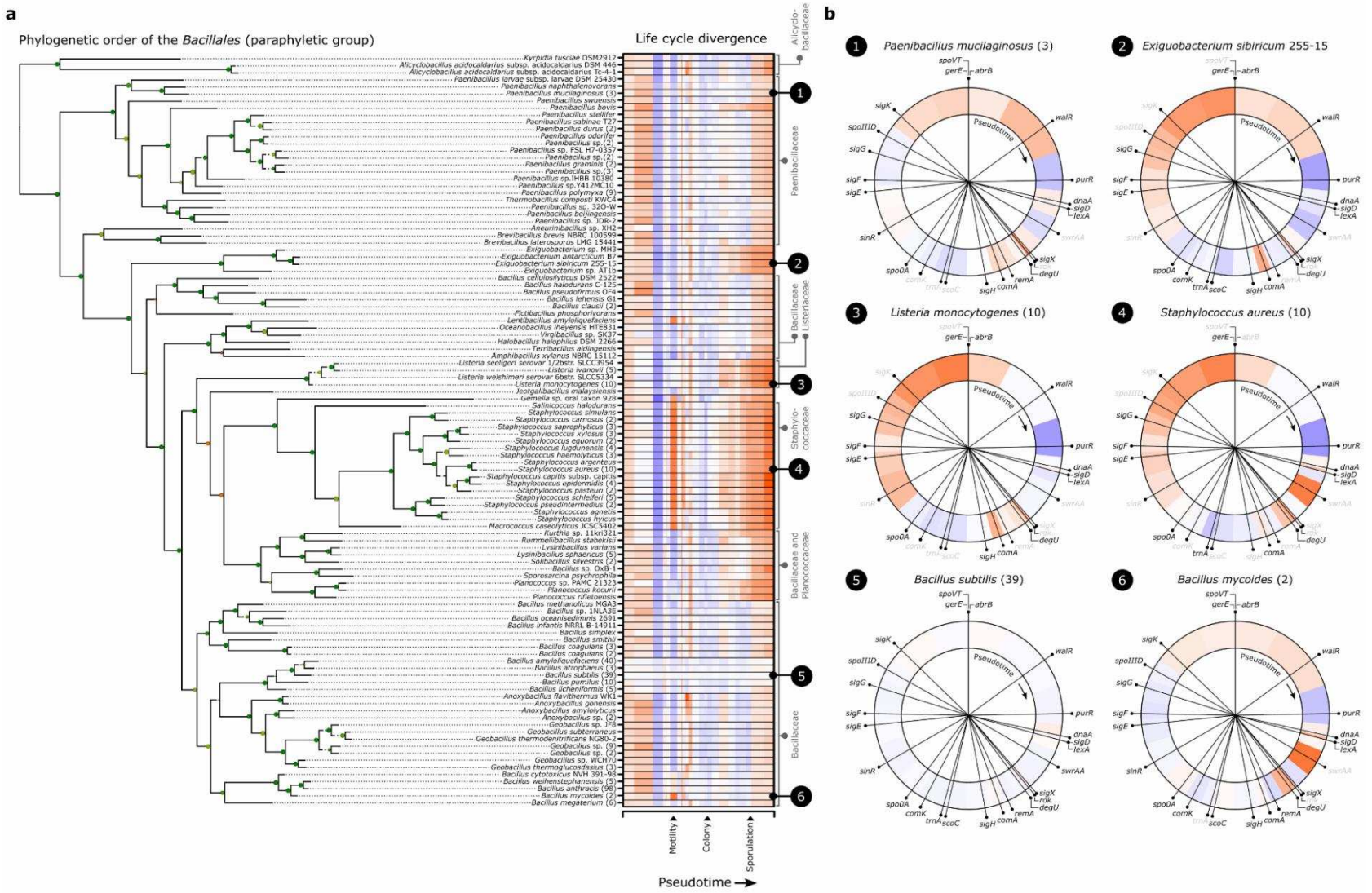
478

479 **Supplementary Fig. 25. Conservation of genes underlying lifestyle switches in *B. subtilis*.** Average conservation of genes across branches of

480 phylogenetic tree (Supplementary Fig. 21 and Supplementary Text 4) based on bi-directional best BLAST hits (e-value < 10^{-10} and coverage >

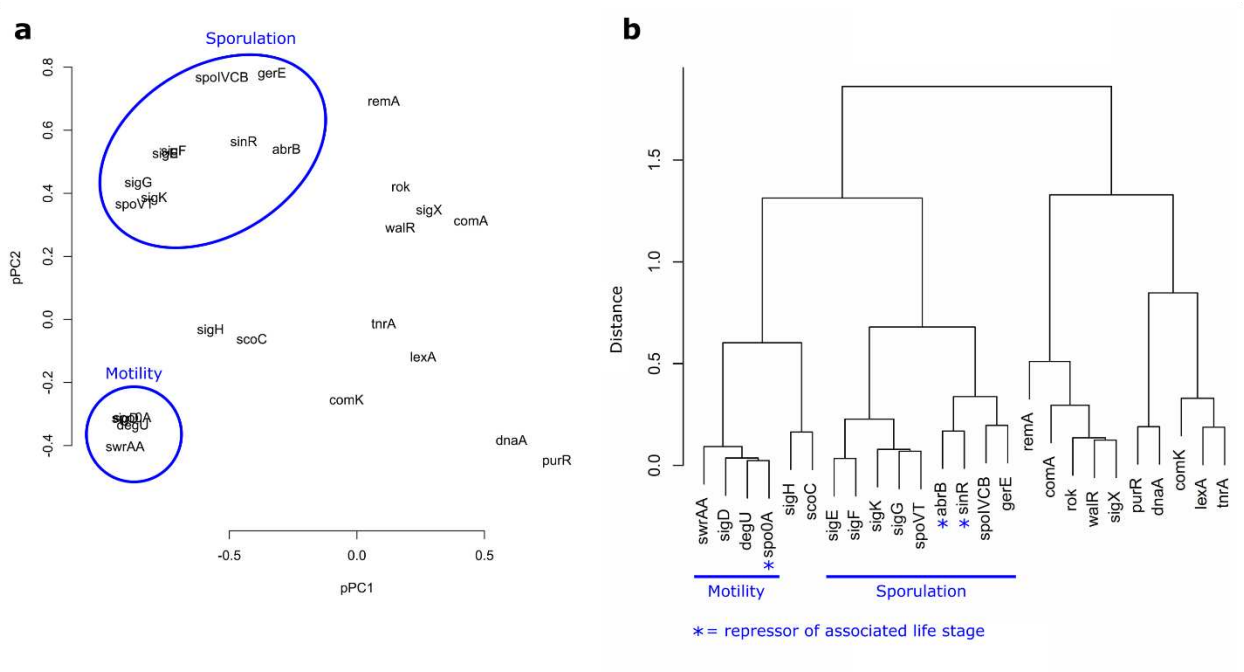
481 70%): red, no conservation of genes; green, strong conservation of genes. For simplicity, genes and proteins have the same color

482

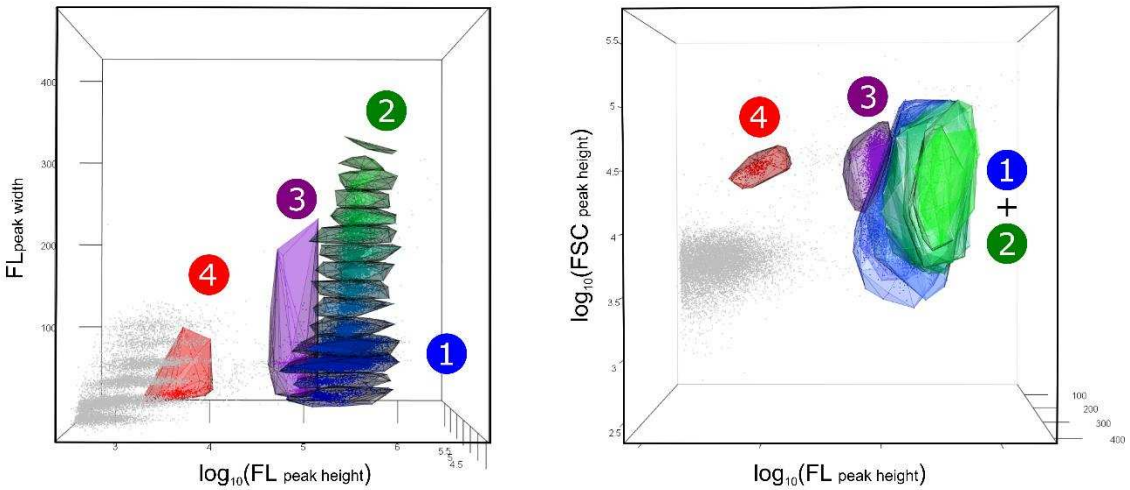


483 **Supplementary Fig S26**. **Mosaic conservation patterns across the phylogenetic tree as shown in Fig. 5, including both orthologs and paralogs**.

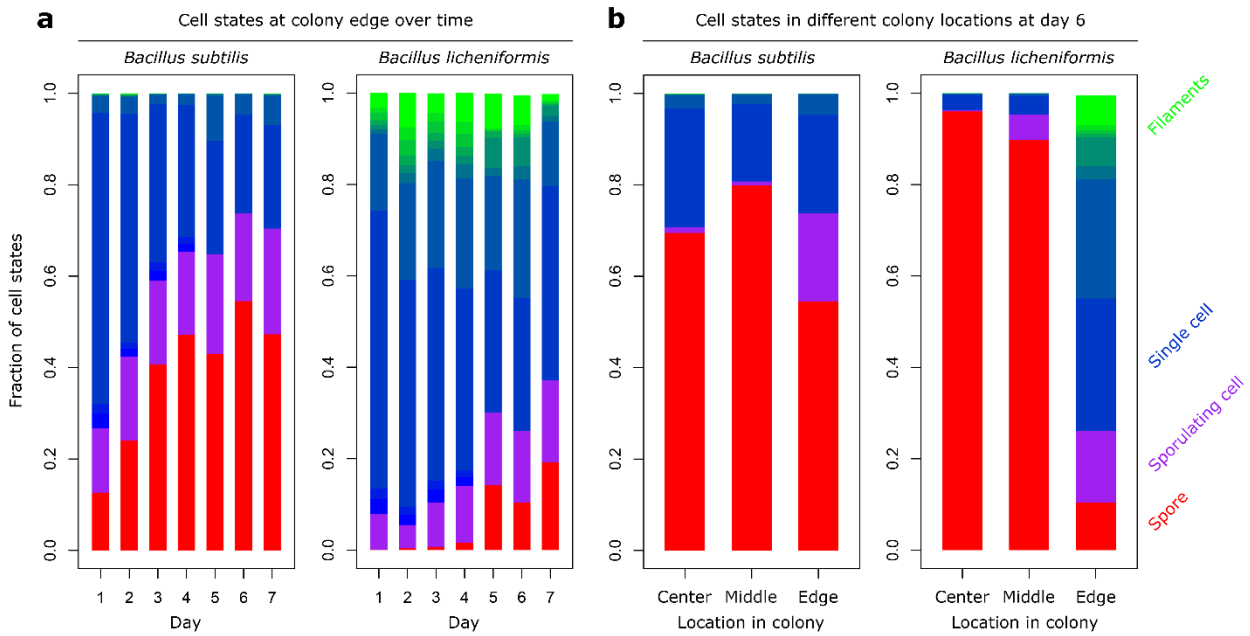
484 In other words, colors represent gene conservation based on unidirectional BLAST hits of genes from the reference genome of *B. subtilis*.



485 **Supplementary Fig. 27. Modularity in regulon conservation patterns**. **a**, Phylogenetic principal
486 component analysis (pPCA) on the conservation of life cycle regulons (Fig. 4b) across the *Bacillales*.
487 Regulons are represented by the name of their associated upstream regulators. Blue ellipses highlight
488 the regulons underlying the motility and sporulation life stages. **b**, Hierarchical clustering algorithm on
489 pPCA shows modular conservation patterns, in which the regulons underlying motility or sporulation
490 express similar conservation patterns across the *Bacillales*.



Supplementary Fig. 28. Side and top view of flow cytometry results shown in Fig. 6b.



494

495 **Supplementary Fig. 29. Cell state composition in *B. subtilis* and *B. licheniformis* colonies**. **a**, Temporal

496 changes in cell state composition at the colony edge over the course of a week: red, spore; purple,

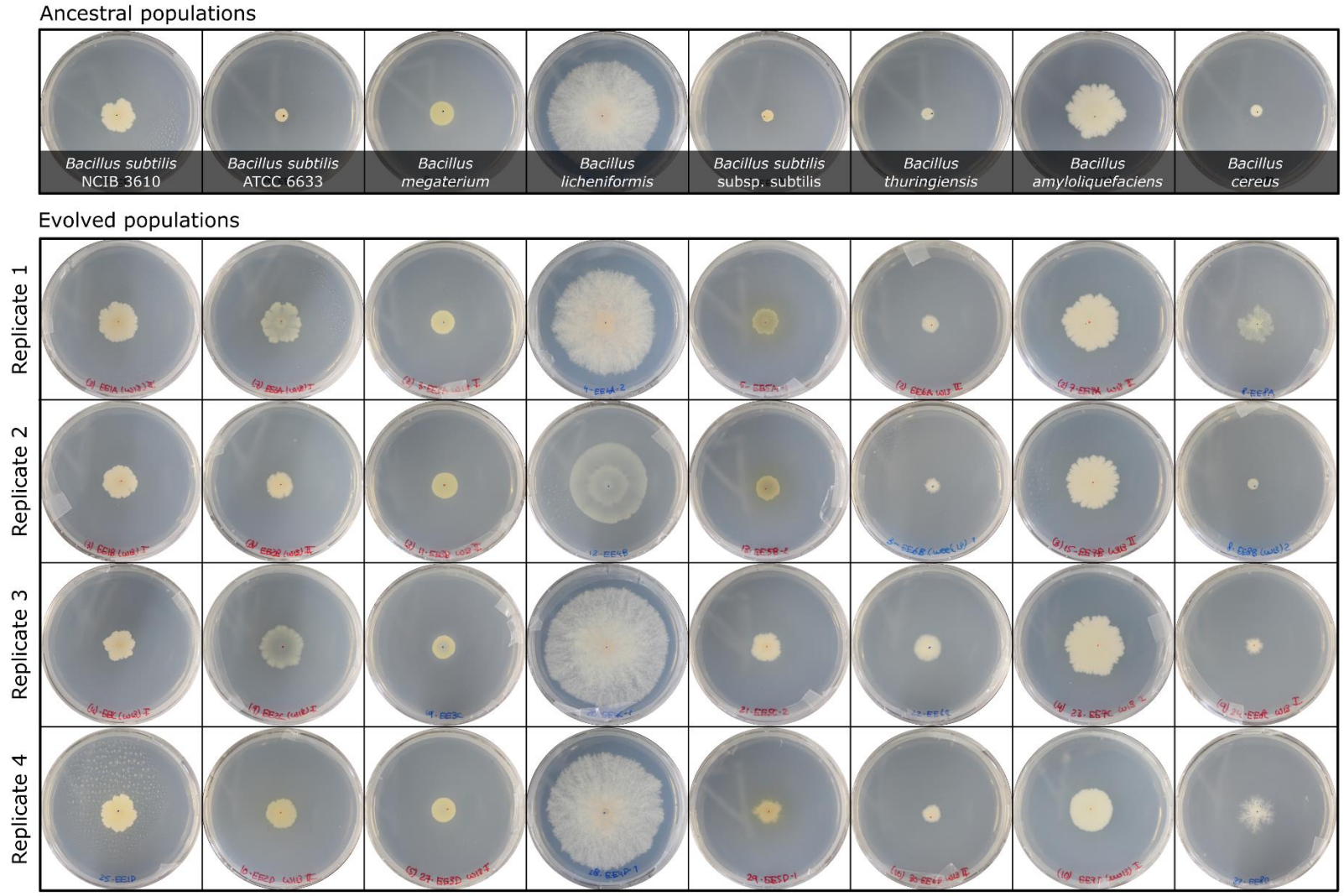
497 sporulating cell; blue, single cell; blue-green, filamentous cells of increasing length (see Methods for

498 applied gating). For simplicity, germinating spores are ignored. **b**, Cell state composition at different

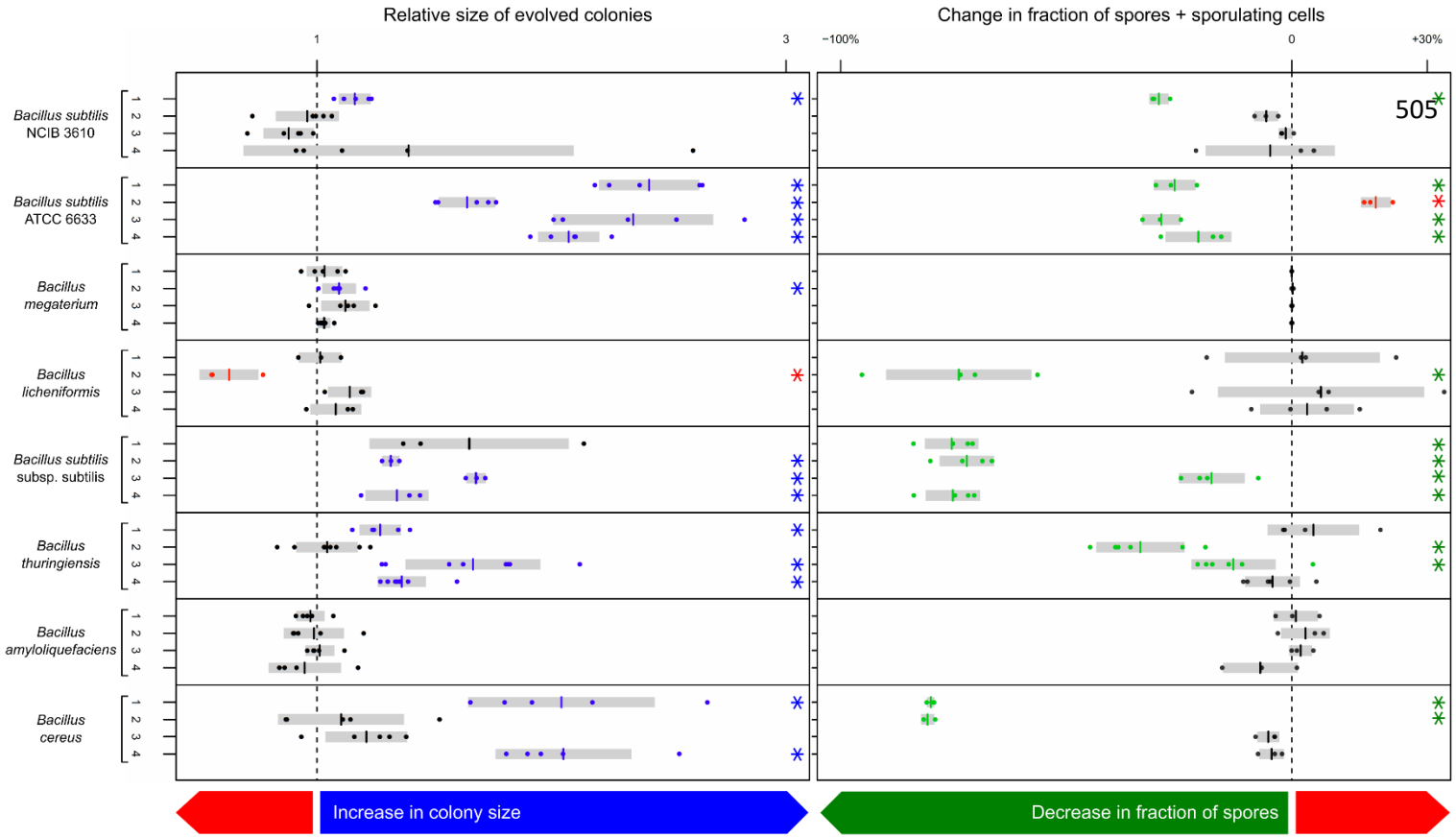
499 locations inside the colony at day 6: center, middle (i.e. halfway from the center to the edge) and edge.

500 Whereas *B. licheniformis* shows a sharp transition in the fraction of spores from the colony center to the

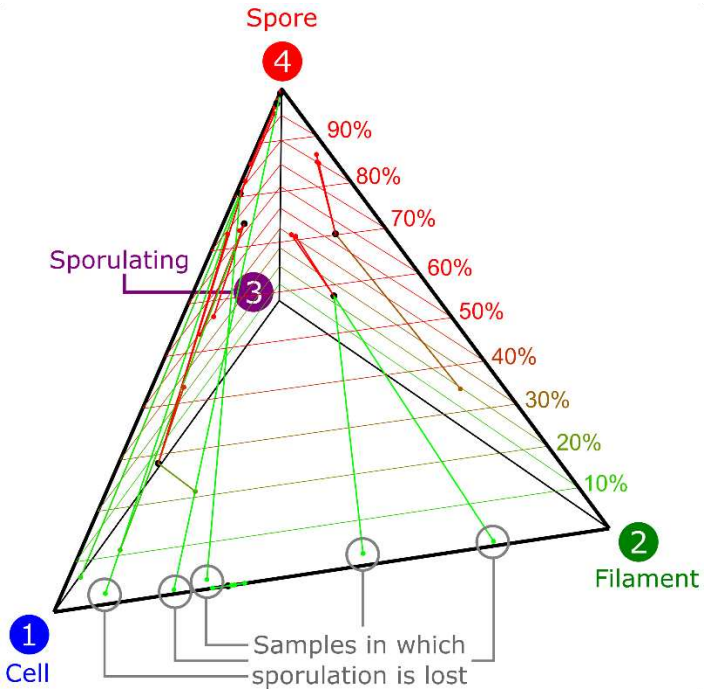
501 edge, *B. subtilis* shows a high fraction of single cells and spores in all locations of the colony.



Supplementary Fig. 30. Evolved colonies. Top row shows representative colonies of the ancestral *Bacillales* strains and species. Bottom rows show representative colonies for the evolved populations.



506 **Supplementary Fig. 31. Evolutionary changes in colony size and cell state composition.** For each of the evolved replicate populations, we show
507 the relative colony size with respect to the ancestor and the change in the fraction of spores and sporulation cells with respect to the ancestor.
508 Relative colony sizes that are significantly smaller/greater than 1 are shown in red/blue. Changes in the fraction of spores and sporulation cells
509 that are significantly lower/higher than 0% are shown in green/red. Asterisks show which populations have a significant change in colony size
510 (one-sample t test, test mean = 1, $P < 0.05$, n = 3-7) and/or cell state composition (one sample t-test, test mean = 0, $P < 0.05$, n = 3-6).



511

512

513 **Supplementary Fig. 32. Tetrahedron of evolutionary changes in cell state composition at colony**

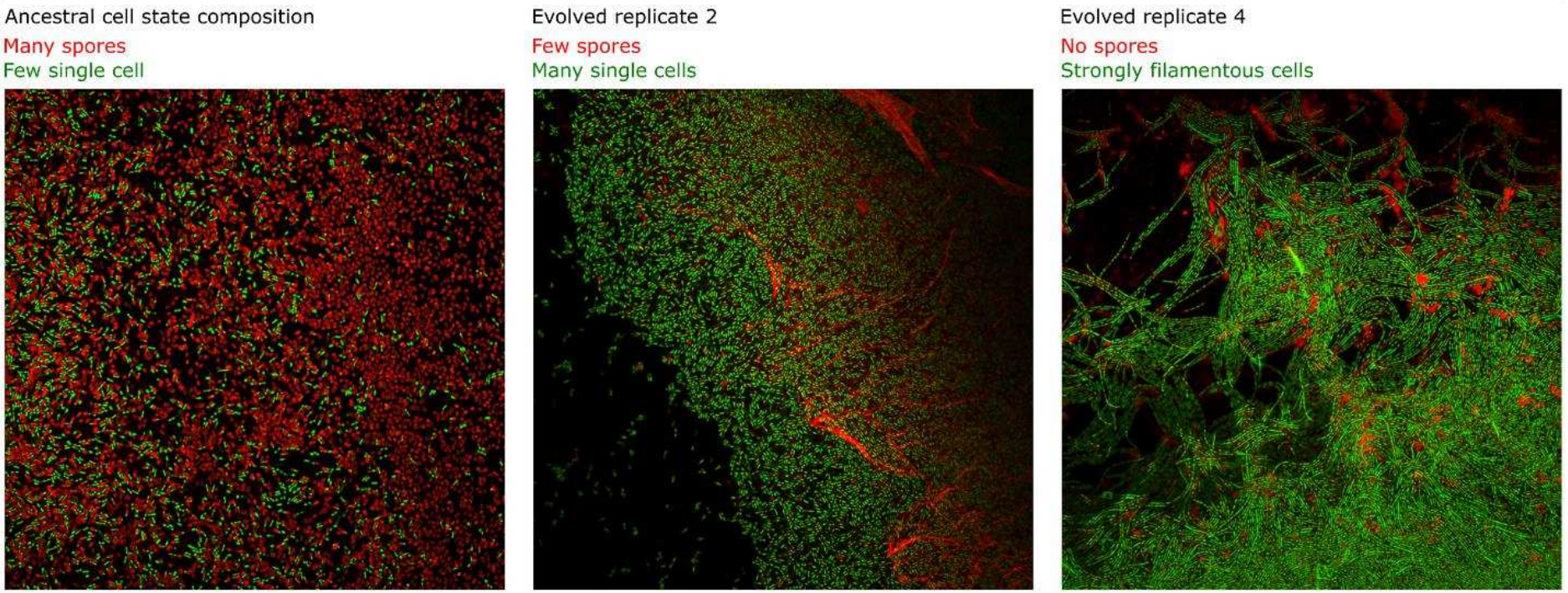
514 **center**. Ancestors are shown by black dots and lines show change in cell state composition in evolved

515 populations. Line colors show the fraction of spores present in the evolved populations: red, 100%

516 spores; green, 0% spores. Grey circles highlight five samples that show complete loss of sporulation at

517 both colony center and colony edge. These samples do not all touch the edge of the tetrahedron, due to

518 data scatter in the flow cytometer analysis.

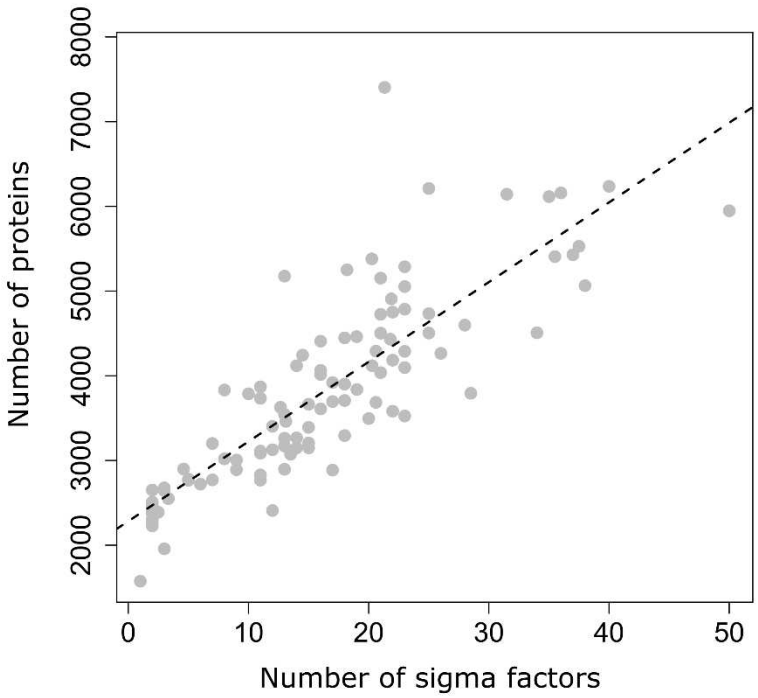


519

520 **Supplementary Fig. 33. Confocal scanning-laser microscopy images of colony edge of *Bacillus subtilis* subsp. subtilis.** Green, SYBR Green

521 stained DNA. Red, MitoTracker Deep Red FM stained cell membranes. Spores can be recognized by the presence of red membrane stain and the

522 absence of green DNA stain. As a by-product, MitoTracker also stained some debris present in the right two images.



Supplementary Fig. 34. Relation between number of sigma factors and genome size. Relation between the average genome size, i.e. number of protein-coding genes, and the number of sigma factors across the phylogenetic tree of the *Bacillales* (Fig. 5a and Supplementary Fig. 21). Dashed line shows linear regression ($R^2 = 0.71$, $P < 10^{-15}$). Number of sigma factors is determined by the number of unique protein BLAST hits (e-value < 10^{-10} and coverage > 70%) with respect to the sigma factors of *B. subtilis*, as listed in the SubtiWiki database[4].

Supplementary Table S4. Examples of taxonomic groups in the Bacteria with life cycles that have specialized cell types, corresponding to motility and/or dormancy, which would be particularly suitable for a neural network analysis as presented in the main text[59].

Phyla	Order	Example genera	Relevant cell types	Description (References)
Cyanobacteria	Pleurocapsales	Xenococcus Dermacarpella Stanieria Myxosarcina Pleurocapsa	Baeocytes	Small spherical cells that typically form fixed starting points of life cycles, which can either be motile or non-motile.
	Nostocales Stigonematales	Anabaena Nostoc Fischerella	Heterocysts	Cells specialized in nitrogen fixation.
			Hormogonia	Small filamentous cells that typically disperse through gliding motility.
			Akinetes	Cells specialized in long-term dormancy, for example resistant to cold and desiccation.
Proteobacteria	Pseudomonadales	Azotobacter	Cysts	Dormant cells.
Proteobacteria	Rhodospirillales	Azospirillum	Cysts	Dormant cells.
Proteobacteria	Myxococcales	Aggregicoccus Myxococcus Stigmatella Melittangium	Swarming cells	Cells that engage in collective gliding motility.
			Fruiting bodies	Cell aggregates that form in response to starvation to mediate spore dispersal.
			Myxospores	Dormant spores that resist for example cold and desiccation.
Actinobacteria	Micromonosporales	Actinoplanes Catenuloplanes Planomonospora Ineosporia Spirillospora	Mycelia	Hyphal growth structures that grow through tip extension and branching.
			Sporangia	Multicellular structures that arise from the mycelia and develop sacs of spores to facilitate dispersal.
			Zoospores	Motile spores that can disperse by swimming and germinate when finding prosperous conditions.
	Streptomycetales	Streptomyces Kitasatospora	Mycelia	Hyphal growth structures that grow through tip extension and branching.
			Aerial hyphae	Specialized hyphae that extend into the air from the mycelium and facilitate dispersal by developing spores at the tips.
			Exospores	Resistant spores.
Firmicutes	Bacillales Clostridiales	Candidatus Bacillus Paenibacillus Clostridium Metabacterium Candidatus	Filaments	Filamentous cells that are attached at the pole-to-pole ends.
			Endospores	Resistant spores that can survive extreme environmental stresses, such as desiccation, heat and cold.
			Twin spores	Alternative form of endosporulation that yields two spores per mother cell. This is just one of the many alternative phenotypic outcomes of endosporulation.

Supplementary Table 7. Studies corroborating evolutionary losses of life stages

Taxonomic group (genera or species)	Life stage absent	Studies confirming the absence of life stage(s)*
Paenibacillus bovis	Sporulation	None**.
Exiguobacterium	Sporulation	Farrow *et al.*, 1994[60]
Lentibacillus amyloliquefaciens	Motility	Wang *et al.,* 2016a[61] Wang *et al.*, 2016b[62]
Listeria	Sporulation	Stragier, 2002[63] Onyenwoke *et al.*, 2004[64] Gründling *et al.*, 2004[65] *** O'Neil & Marquis, 2006[66] *** Abecasis *et al.*, 2013[67] Galperin 2013[68]
Jeotgalibacillus malaysiensis	No loss****	Yaakop *et al*. 2015[69] Goh *et al.*, 2015[70]
Gemella *Salinicoccus* *Staphylococcus* *Macrococcus*	Motility and sporulation	Kloos *et al.*, 1998[71] Stragier, 2002[63] Onyenwoke *et al.*, 2004[64] Wang *et al.*, 2008[72] Baba *et al.,* 2009[73] Abecasis *et al.*, 2013[67] Galperin 2013[68] Jiang *et a.*, 2015[74]
Planococcus	Sporulation	Farrow *et al.*, 1994[60]
Bacillus mycoides	Motility	Nakamura & Jackson, 1995[75] Di Franco *et al*., 2002[76] Soufiane & Côté, 2013[77]

* Selection of studies (and references therein) that confirm the lack of the motility and/or sporulation life stage(s) through either genomic or phenotypic analyses. Some of this evidence is anecdotal. Importantly, the absence of a specific life stage in standard phenotypic assays is not strict evidence for the loss of this stage, because the phenotypes of some species may not be expressed under standard laboratory condition. For example, *Bacillus anthracis* was long thought to be strictly non-motile, whereas motile variants do exist. By the same token, many *Paenibacilli* do not sporulate in standard laboratory settings, even when most genes underlying sporulation are conserved.

** In contradiction to our bioinformatic analysis, Gao *et al*. (2016)[78] claim that *Paenibacillus bovis* sporulates. Their evidence consists of a few inconclusive microscopy images. We therefore suggest that this species needs to be characterized in further detail.

*** Confirmation that the motile life stage in *Listeria monocytogenes* is conserved, despite weak conservation of motility genes, as apparent from our bioinformatic analysis.

**** Support for the conservation of life stages in *Jetgalibacillus malaysiensis*, which is closely related to the *Listeria* and *Staphylococcus*, two genera that did lose the motility and/or sporulation life stages.

Supplementary Table 8. Strains used for the evolution experiment.

n	strain	Source	RefSeq assembly accession
1	Bacillus subtilis subsp. subtilis NCIB 3610	Kolter lab	GCF_000186085.1
2	Bacillus subtilis subsp. spizizenii ATCC 6633	Kolter lab	GCF_000177595.1
3	Bacillus megaterium ATCC 14581	Kolter lab	GCF_000832985.1
4	Bacillus licheniformis ATCC 14580	BGSC*	GCF_000011645.1
5	Bacillus subtilis subsp. subtilis	JvG	GCF_000971925.1**
6	Bacillus thuringiensis subsp. kurstaki HD73	BGSC	GCF_000338755.1
7	Bacillus amyloliquefaciens subsp. plantarum FZB42	BGSC	GCF_000015785.1
8	Bacillus cereus ATCC 10987	BGSC	GCF_000008005.1

* Bacillus Genetic Stock Center, Biological Sciences 556, 484 W. 12th Ave, Columbus, OH 43210-1214

** This is the closest reference genome, because this is a non-reference strain

554 **Supplementary Table 9**. Overview of mutations in populations that lost sporulation in our evolution experiment

				Clones*					
				Sporulating		Non-sporulating		Provean	
Population	Gene(s)**	Position	Mutation	1	2	3	4	Score[#]	Description / function
B. licheniformis Replicate 2	*TRNA_RS27060-TRNA_RS27065*	1,121,114	A→G		●	●	●	NA	Intergenic mutation
	TRNA_RS35945	2,807,704	S69P (TCA→CCA)			●	●	-3.433	Ortholog of *spo0B* in *B. subtilis*. Spo0B functions as a sporulation initiation phosphotransferase by phosphorylating the global regulator Spo0A.***
	pgsB	3,667,729	W2* (TGG→TAG)		●	●	●	NA	Poly-gamma-glutamate synthase
	TRNA_RS41335	3,857,813	$(ATCTTTACC)_{1\rightarrow2}$			●	●	-11.867	ABC transporter ATP-binding protein
B. subtilis subsp. subtilis Replicate 1	*kinA*	1,470,767	$(CTG)_{4\rightarrow2}$				●	-5.072	Histidine kinase involved in the induction of sporulation, through the phosphorylation cascade that results in the phosphorylation of Spo0A
	artP	2,492,278	Δ5 bp				●	NA	High affinity arginine ABC transporter binding lipoprotein
	spo0F-ywjG	3,809,940	Δ26 bp			●		NA	Deletion in promoter of *spo0F*; Spo0F is the two-component response regulator that – through Spo0B – leads to the phosphorylation of Spo0A.
B. subtilis subsp. subtilis Replicate 4	*abrB*	44,933	L69P (CTT→CCT)		●	●	●	-5.857	Transcriptional regulator for transition state genes
	[yqiI] - [recN]	2,515,444	Δ5,651 bp			●	●	NA	Large genomic deletion that includes the gene, *spo0A*, encoding for the global regulator Spo0A.
	spo0B	2,854,290	M90I (ATG→ATA)			●		0.213	Sporulation initiation phosphotransferase, responsible for phosphorylating the global regulator Spo0A.
	epsG	3,522,340	Y312N (TAT→AAT)		●			-8.583	Biofilm extracellular matrix formation
	[gtaB] - [yvzE]	3,666,058	Δ6,907 bp		●	●	●	NA	Deletion of number of genes, including genes underlying poly(glucosyl N-acetylgalactosamine 1-phosphate) glucosyltransferase
B. cereus Replicate 1	*BCE_RS00275*	50,532	2,271 bp → IS4-like element			●	●	NA	IS4-like element ISBce2 family transposase integrated at promoter of *spoVG*; SpoVG is repressor of asymmetrical septum formation during sporulation[79,80]
B. cereus Replicate 2	*BCE_RS01280*	236,214	Q96* (CAA→TAA)				●	NA	LacI family DNA-binding transcriptional regulator
	spo0A	3,956,556	T98I (ACA→ATA)			●	●	-4.822	Global regulator of sporulation

555 * Clone 1 = ancestral clone; clone 2 = clone before loss of sporulation; clone 3 & 4 = two clones (right) after loss of sporulation.

556 ** Accessions of reference genomes used to map reads: *B. licheniformis* = NC_006270; *B. subtilis* = NC_000964; *B. cereus* = NC_003909 & NC_005707.

557 *** Mutations that potentially resulted in loss of sporulation have gray background. Except for the mutation in *BCE_RS00275*, all these mutations affect Spo0A.

558 [#] Provean scores assess functional consequences of amin oacid subsitutions/indels[81,82], Provean score < -2.5 are considered deleterious to protein function.

559 **Supplementary Table 10. Medium components.**

#	Chemical compound	Chemical Formula	Concentration	Company
1	Sodium Phosphate Buffer (pH 7.3)	Na_2HPO_4 + NaH_2PO_4	2.5mM	Fluka #71645, Fluka #71500
2	Glycerol	$C_3H_8O_3$	5.4mM (=0.05%)	AppliChem #A3739
3	Glutamate	$C_5H_8NNaO_4$	30mM (=0.5%)	Sigma-Aldrich #G5889
4	MOPS (pH 7)	$C_7H_{15}NO_4S$	0.01M	AppliChem #A2947
5	Manganese(II) chloride	$MnCl_2$	0.05mM	Sigma-Aldrich #31422
6	Magnesiumchloride	$MgCl_2$	2mM	Fluka #63065
7	Ferric chloride	$FeCl_3$	0.05mM	Sigma-Aldrich #236489
8	Calciumchloride	$CaCl_2$	0.7mM	Sigma-Aldrich #C3881
9	Potassium chloride	KCl	0.1mM	Sigma-Aldrich #60130
10	Zinc chloride	$ZnCl_2$	1uM	Merck #7815

560

561 **Supplementary Table 11. PCR primer sets for checking potential cross-contamination between**

562 **species/strains.**

n	Forward primer	Backward primer	Target gene
1	CCCTACCGCTTTTTGTTTTGC	AATTGCTGTTCCTTCTGCCC	[locus_tag=B4U62_RS18230] [protein=sporulation-delaying protein SdpB]
2	CATACTTTCCTGCAAGCGCC	TCTGAAACAGTCATCGGCGG	[locus_tag=BSUW23_RS16860] [protein=SpaB]
3	GAGCAAAAAGCCGTTCCAGG	ACTGCTGTCCACTTTACGGG	[locus_tag=BG04_RS22120] [protein=betagalactosidase]
4	AACTCCCTGCTGCTTCATCC	TGTGGGAATCCAGTTCCTGC	[locus_tag=TRNA_RS39345] [protein=primase]
5	CCCTACCGCTTTTTGTTTTGC	AATTGCTGTTCCTTCTGCCC	[locus_tag=O7A_RS18340] [protein=sporulation-delaying protein SdpB]
6	TCTGGAAAGAGCGCAGAAGG	ATTCCCTAGTCTTGCGTGCC	[locus_tag=HD73_RS29455] [protein=pesticidal crystal protein cry1Ac]
7	TAACGATGAGCCATGTCGGG	CGCCGCATCCATTATCTTGC	[locus_tag=RBAM_RS10975] [protein=D-fructose-6-phosphate amidotransferase]
8	TTGTAGTTCCGAGTCCTGCG	TCTCTTGGGAAACGATGCGG	[locus_tag=BCE_RS05305] [protein=type III restriction-modification system Bce10987IP enzyme res]

563 * For the very closely-related strains 1 and 5 the same primer sets were used

References

1. Sierro, N., Makita, Y., de Hoon, M. & Nakai, K. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res.* **36**, D93-96 (2008).
2. Freyre-González, J. A. *et al.* Lessons from the modular organization of the transcriptional regulatory network of *Bacillus subtilis*. *BMC Syst. Biol.* **7**, 127 (2013).
3. Leyn, S. A. *et al.* Genomic reconstruction of the transcriptional regulatory network in *Bacillus subtilis*. *J. Bacteriol.* **195**, 2463–2473 (2013).
4. Michna, R. H., Zhu, B., Mäder, U. & Stülke, J. *Subti*Wiki 2.0-an integrated database for the model organism *Bacillus subtilis*. *Nucleic Acids Res.* **44**, D654-662 (2016).
5. Bonneau, R. *et al.* The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biol.* **7**, R36 (2006).
6. Fadda, A. *et al.* Inferring the transcriptional network of *Bacillus subtilis*. *Mol. Biosyst.* **5**, 1840–1852 (2009).
7. Arrieta-Ortiz, M. L. *et al.* An experimentally supported model of the *Bacillus subtilis* global transcriptional regulatory network. *Mol. Syst. Biol.* **11**, 839 (2015).
8. Flórez, L. A., Roppel, S. F., Schmeisky, A. G., Lammers, C. R. & Stülke, J. A community-curated consensual annotation that is continuously updated: the *Bacillus subtilis* centred wiki *Subti*Wiki. *Database* 1–9 (2009).
9. Mäder, U., Schmeisky, A. G., Flórez, L. A. & Stülke, J. *Subti*Wiki—a comprehensive community resource for the model organism *Bacillus subtilis*. *Nucleic Acids Res.* **40**, D1278–D1287 (2012).
10. Michna, R. H., Commichau, F. M., Tödter, D., Zschiedrich, C. P. & Stülke, J. *Subti*Wiki–a database for the model organism *Bacillus subtilis* that links pathway, interaction and expression information. *Nucleic Acids Res.* **42**, D692–D698 (2014).

11. Zhu, B. & Stülke, J. *Subti*Wiki in 2018: from genes and proteins to functional network annotation of the model organism *Bacillus subtilis*. *Nucleic Acids Res.* **46**, D743–D748 (2018).
12. Kramer, M. A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* **37**, 233–243 (1991).
13. Scholz, M., Fraunholz, M. & Selbig, J. Nonlinear principal component analysis: neural network models and applications. in *Principal Manifolds for Data Visualization and Dimension Reduction* 44–67 (Springer, Berlin, Heidelberg, 2008).
14. Scholz, M. Validation of nonlinear PCA. *Neural Process. Lett.* **36**, 21–30 (2012).
15. Scholz, M. & Fraunholz, M. J. A computational model of gene expression reveals early transcriptional events at the subtelomeric regions of the malaria parasite, *Plasmodium falciparum*. *Genome Biol.* **9**, R88 (2008).
16. Nicolas, P. *et al.* Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science* **335**, 1103–1106 (2012).
17. Norman, T. M., Lord, N. D., Paulsson, J. & Losick, R. Stochastic switching of cell fate in microbes. *Annu. Rev. Microbiol.* **69**, 381–403 (2015).
18. Russell, J. R., Cabeen, M. T., Wiggins, P. A., Paulsson, J. & Losick, R. Noise in a phosphorelay drives stochastic entry into sporulation in *Bacillus subtilis*. *EMBO J.* e201796988 (2017).
19. Norman, T. M., Lord, N. D., Paulsson, J. & Losick, R. Memory and modularity in cell-fate decision making. *Nature* **503**, 481–486 (2013).
20. Dubnau, D. & Losick, R. Bistability in bacteria. *Mol. Microbiol.* **61**, 564–572 (2006).
21. Veening, J. W., Smits, W. K. & Kuipers, O. P. Bistability, epigenetics, and bet-hedging in bacteria. *Annu. Rev. Microbiol.* **62**, 193–210 (2008).
22. Lopez, D., Vlamakis, H. & Kolter, R. Generation of multiple cell types in *Bacillus subtilis*. *FEMS Microbiol. Rev.* **33**, 152–163 (2009).

612 23. van Gestel, J., Vlamakis, H. & Kolter, R. From cell differentiation to cell collectives: *Bacillus subtilis*
613 uses division of labor to migrate. *PLoS Biol* **13**, e1002141 (2015).
614 24. Stragier, P. & Losick, R. Molecular genetics of sporulation in *Bacillus subtilis*. *Annu. Rev. Genet.* **30**,
615 297–341 (1996).
616 25. Kroos, L. The *Bacillus* and *Myxococcus* developmental networks and their transcriptional regulators.
617 *Annu. Rev. Genet.* **41**, 13–39 (2007).
618 26. Higgins, D. & Dworkin, J. Recent progress in *Bacillus subtilis* sporulation. *FEMS Microbiol. Rev.* **36**,
619 131–148 (2012).
620 27. Haldenwang, W. G. The sigma factors of *Bacillus subtilis*. *Microbiol. Rev.* **59**, 1–30 (1995).
621 28. Errington, J. & Mandelstam, J. Use of a *lacZ* gene fusion to determine the dependence pattern of
622 sporulation operon *spoIIA* in *spo* mutants of *Bacillus subtilis*. *Microbiology* **132**, 2967–2976 (1986).
623 29. Satola, S. W., Baldus, J. M. & Moran, C. P. Binding of Spo0A stimulates *spoIIG* promoter activity in
624 *Bacillus subtilis*. *J. Bacteriol.* **174**, 1448–1453 (1992).
625 30. Bird, T. H., Grimsiey, J. K., Hoch, J. A. & Spiegelman, G. B. Phosphorylation of Spo0A activates its
626 stimulation of *in vitro* transcription from the *Bacillus subtilis spoIIG* operon. *Mol. Microbiol.* **9**, 741–
627 749 (1993).
628 31. Dworkin, J. & Losick, R. Developmental commitment in a bacterium. *Cell* **121**, 401–409 (2005).
629 32. Partridge, S. R., Foulger, D. & Errington, J. The role of σ^F in prespore-specific transcription in *Bacillus*
630 *subtilis*. *Mol. Microbiol.* **5**, 757–767 (1991).
631 33. Karow, M. L., Glaser, P. & Piggot, P. J. Identification of a gene, *spoIIR*, that links the activation of σ^E
632 to the transcriptional activity of σ^F during sporulation in *Bacillus subtilis*. *Proc. Natl. Acad. Sci. U. S.*
633 *A.* **92**, 2012–2016 (1995).

634 34. Meisner, J., Wang, X., Serrano, M., Henriques, A. O. & Moran, C. P. A channel connecting the mother
635 cell and forespore during bacterial endospore formation. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 15100–
636 15105 (2008).
637 35. Wang, S. T. *et al.* The forespore line of gene expression in *Bacillus subtilis*. *J. Mol. Biol.* **358**, 16–37
638 (2006).
639 36. Eichenberger, P. *et al.* The program of gene transcription for a single differentiating cell type during
640 sporulation in *Bacillus subtilis*. *PLoS Biol.* **2**, e328 (2004).
641 37. Kunkel, B., Sandman, K., Panzer, S., Youngman, P. & Losick, R. The promoter for a sporulation gene
642 in the *spoIVC* locus of *Bacillus subtilis* and its use in studies of temporal and spatial control of gene
643 expression. *J. Bacteriol.* **170**, 3513–3522 (1988).
644 38. Cutting, S., Driks, A., Schmidt, R., Kunkel, B. & Losick, R. Forespore-specific transcription of a gene in
645 the signal transduction pathway that governs Pro-σ^K processing in *Bacillus subtilis*. *Genes Dev.* **5**,
646 456–466 (1991).
647 39. Dong, T. C. & Cutting, S. M. SpoIVB-mediated cleavage of SpoIVFA could provide the intercellular
648 signal to activate processing of Pro-σ^K in *Bacillus subtilis*. *Mol. Microbiol.* **49**, 1425–1434 (2003).
649 40. Campo, N. & Rudner, D. Z. SpoIVB and CtpB are both forespore signals in the activation of the
650 sporulation transcription factor σK in *Bacillus subtilis*. *J. Bacteriol.* **189**, 6021–6027 (2007).
651 41. Zheng, L. & Losick, R. Cascade regulation of spore coat gene expression in *Bacillus subtilis*. *J. Mol.*
652 *Biol.* **212**, 645–660 (1990).
653 42. Nugroho, F. A., Yamamoto, H., Kobayashi, Y. & Sekiguchi, J. Characterization of a new sigma-K-
654 dependent peptidoglycan hydrolase gene that plays a role in *Bacillus subtilis* mother cell lysis. *J.*
655 *Bacteriol.* **181**, 6230–6237 (1999).
656 43. Mielich-Süss, B. & Lopez, D. Molecular mechanisms involved in *Bacillus subtilis* biofilm formation.
657 *Environ. Microbiol.* **17**, 555–565 (2015).

44. Smits, W. K. *et al.* Stripping *Bacillus*: ComK auto-stimulation is responsible for the bistable response in competence development. *Mol. Microbiol.* **56**, 604–614 (2005).
45. Fujita, M., González-Pastor, J. E. & Losick, R. High- and low-threshold genes in the Spo0A regulon of *Bacillus subtilis*. *J. Bacteriol.* **187**, 1357–1368 (2005).
46. Grau, R. R. *et al.* A duo of potassium-responsive histidine kinases govern the multicellular destiny of *Bacillus subtilis*. *mBio* **6**, e00581 (2015).
47. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of *Bacteria* and *Archaea*. *Nature* **462**, 1056–1060 (2009).
48. Wu, D., Jospin, G. & Eisen, J. A. Systematic identification of gene families for use as 'markers' for phylogenetic and phylogeny-driven ecological studies of *Bacteria* and *Archaea* and their major subgroups. *PLoS One* **8**, e77033 (2013).
49. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
50. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
51. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
52. Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10557–10562 (2005).
53. Löytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632–1635 (2008).
54. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).

55. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007).
56. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
57. Zhang, J. & Yang, J. R. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* **16**, 409–420 (2015).
58. Bergman, N. H. *et al.* Transcriptional profiling of the *Bacillus anthracis* life cycle *in vitro* and an implied model for regulation of spore formation. *J. Bacteriol.* **188**, 6092–6100 (2006).
59. Shapiro, J. A. Bacteria as multicellular organisms. *Sci. Am.* **258**, 82–89 (1988).
60. Farrow, J. A., Wallbanks, S. & Collins, M. D. Phylogenetic interrelationships of round-spore-forming Bacilli containing cell walls based on lysine and the non-spore-forming genera *Caryophanon*, *Exiguobacterium*, *Kurthia*, and *Planococcus*. *Int. J. Syst. Bacteriol.* **44**, 74–82 (1994).
61. Wang, J. L. *et al.* Complete genome sequence of strain *Lentibacillus amyloliquefaciens* LAM0015T isolated from saline sediment. *J. Biotechnol.* **220**, 88–89 (2016).
62. Wang, J. L. *et al. Lentibacillus amyloliquefaciens* sp. nov., a halophilic bacterium isolated from saline sediment sample. *Antonie van Leeuwenhoek* **109**, 171–178 (2016).
63. Stragier, P. A Gene Odyssey: Exploring the Genomes of Endospore-Forming Bacteria. in *Bacillus subtilis and Its Closest Relatives* (eds. Sonenshein, A. L., Hoch, J. A. & Losick, R.) 519–525 (ASM Press, 2002).
64. Onyenwoke, R. U., Brill, J. A., Farahi, K. & Wiegel, J. Sporulation genes in members of the low G+C Gram-type-positive phylogenetic branch (Firmicutes). *Arch. Microbiol.* **182**, 182–192 (2004).
65. Gründling, A., Burrack, L. S., Bouwer, H. G. A. & Higgins, D. E. *Listeria monocytogenes* regulates flagellar motility gene expression through MogR, a transcriptional repressor required for virulence. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12318–12323 (2004).

66. O'Neil, H. S. & Marquis, H. *Listeria monocytogenes* flagella are used for motility, not as adhesins, to increase host cell invasion. *Infect. Immun.* **74**, 6675–6681 (2006).
67. Abecasis, A. B. *et al.* A genomic signature and the identification of new sporulation genes. *J. Bacteriol.* **195**, 2101–2115 (2013).
68. Galperin, M. Y. Genome diversity of spore-forming Firmicutes. *Microbiol. Spectr.* **1**, 1–15 (2013).
69. Yaakop, A. S. *et al.* Isolation of *Jeotgalibacillus malaysiensis* sp. nov. from a sandy beach, and emended description of the genus *Jeotgalibacillus*. *Int. J. Syst. Evol. Microbiol.* **65**, 2215–2221 (2015).
70. Goh, K. M., Chan, K.-G., Yaakop, A. S. & Ee, R. Complete genome of *Jeotgalibacillus malaysiensis* D5T consisting of a chromosome and a circular megaplasmid. *J. Biotechnol.* **204**, 13–14 (2015).
71. Kloos, W. E. *et al.* Delimiting the genus *Staphylococcus* through description of *Macrococcus caseolyticus* gen. nov., comb. nov. and *Macrococcus equipercicus* sp. nov., and *Macrococcus bovicus* sp. no. and *Macrococcus carouselicus* sp. nov. *Int. J. Syst. Bacteriol.* **48**, 859–877 (1998).
72. Wang, X., Xue, Y., Yuan, S., Zhou, C. & Ma, Y. *Salinicoccus halodurans* sp. nov., a moderate halophile from saline soil in China. *Int. J. Syst. Evol. Microbiol.* **58**, 1537–1541 (2008).
73. Baba, T. *et al.* Complete genome sequence of *Macrococcus caseolyticus* strain JSCS5402, reflecting the ancestral genome of the human-pathogenic staphylococci. *J. Bacteriol.* **191**, 1180–1190 (2009).
74. Jiang, K., Xue, Y. & Ma, Y. Complete genome sequence of *Salinicoccus halodurans* H3B36, isolated from the Qaidam Basin in China. *Stand. Genomic Sci.* **10**, (2015).
75. Nakamura, L. K. & Jackson, M. A. Clarification of the taxonomy of *Bacillus mycoides*. *Int. J. Syst. Evol. Microbiol.* **45**, 46–49 (1995).
76. Di Franco, C., Beccari, E., Santini, T., Pisaneschi, G. & Tecce, G. Colony shape as a genetic trait in the pattern-forming *Bacillus mycoides*. *BMC Microbiol.* **2**, 33 (2002).

77. Soufiane, B. & Côté, J. C. *Bacillus weihenstephanensis* characteristics are present in *Bacillus cereus* and *Bacillus mycoides* strains. *FEMS Microbiol. Lett.* **341**, 127–137 (2013).
78. Gao, C. *et al. Paenibacillus bovis* sp. nov., isolated from raw yak (*Bos grunniens*) milk. *Int. J. Syst. Evol. Microbiol.* **66**, 1413–1418 (2016).
79. Zuber, P. & Losick, R. Role of AbrB in Spo0A- and Spo0B-dependent utilization of a sporulation promoter in *Bacillus subtilis*. *J. Bacteriol.* **169**, 2223–2230 (1987).
80. Matsuno, K. & Sonenshein, A. L. Role of SpoVG in asymmetric septation in *Bacillus subtilis*. *J. Bacteriol.* **181**, 3392–3401 (1999).
81. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).
82. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).